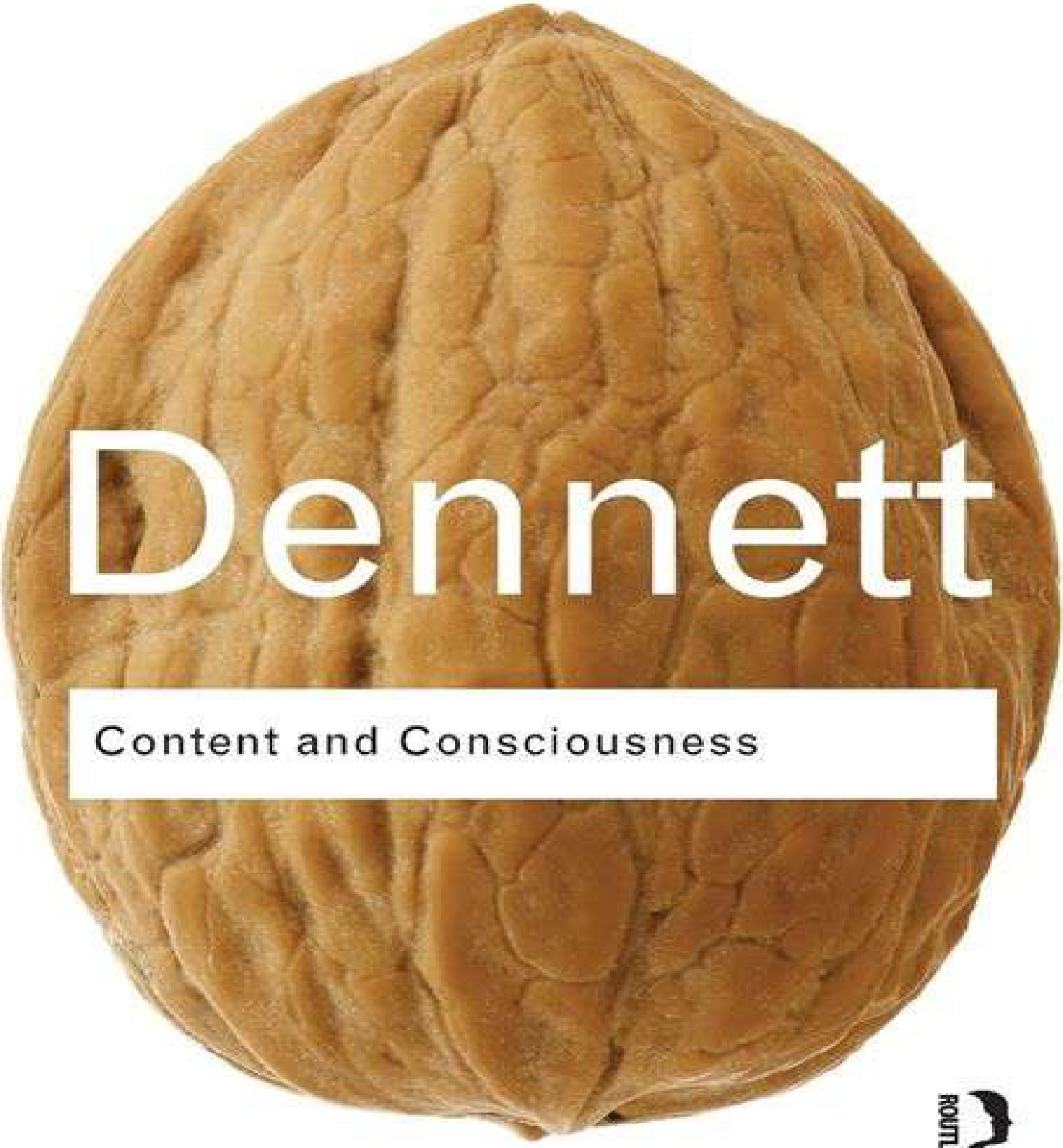


Dennett

Content and Consciousness

ROUTLEDGE

‘ *Content and Consciousness* is an extraordinarily interesting and original book, and one which will raise the level of current discussion in the philosophy of mind.’

Richard Rorty, Philosophical Studies

‘I have certainly been greatly stimulated by reading the book, and I recommend it to all others who have an interest in the problems of

mind and body and of physicalism and its alternatives.’

J.C.C. Smart, Mind

Routledge Classics contains the very best of Routledge Publishing over

the past century or so, books that have, by popular consent, become

established as classics in their field. Drawing on a fantastic heritage of innovative writing published by Routledge and its associated imprints, this series makes available in attractive, affordable form some of the most

important works of modern times.

For a complete list of titles visit

ww.routledge.com/classics

Daniel C. Dennett

Content and Consciousness

With a new preface by the author

London and New York

First published 1969

by Routledge & Kegan Paul

Second edition published by Routledge 1986

First published in Routledge Classics 2010

by Routledge

2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

Simultaneously published in the USA and Canada

by Routledge

270 Madison Avenue, New York, NY 10016

Routledge is an imprint of the Taylor & Francis Group, an informa business
This edition published in the Taylor & Francis e-Library, 2010.

To purchase your own copy of this or any of Taylor & Francis or Routledge's collection of thousands of eBooks please go to www.eBookstore.tandf.co.uk.

© 1969, 1986, 2010 Daniel C. Dennett

All rights reserved. No part of this book may be reprinted

or reproduced or utilised in any form or by any electronic,

mechanical, or other means, now known or hereafter

invented, including photocopying and recording, or in

any information storage or retrieval system, without

permission in writing from the publishers.

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication Data

A catalog record for this book has been requested

ISBN 0-203-09295-3 Master e-book ISBN

ISBN 10: 0–415–56786–6 (pbk)

ISBN 10: 0–203–09295–3 (ebk)

ISBN 13: 978–0–415–56786–2 (pbk)

ISBN 13: 978–0–203–09295–8 (ebk)

For

Susan

CONTENTS

PREFACE TO THE ROUTLEDGE CLASSICS EDITION XI

PREFACE TO SECOND EDITION XV

PREFACE TO FIRST EDITION XIX

PART I The Language of Mind 1

1 The Ontological Problem of Mind 3

I The Mind and Science 3

II Existence and Identity 6

2 Intentionality 20

III The Problem of Intentionality 20

IV Two Blind Alleys 35

V The Way Out 42

3 Evolution in the Brain 46

VI The Intelligent Use of Information 46

VII The Evolution of Appropriate Structures 51

VIII Goal-directed Behaviour 70

4 The Ascription of Content 80

IX Function and Content 80

X Language and Content 92

XI Personal and Sub-personal Levels of Explanation: Pain 101

PART II Consciousness 109

5 Introspective Certainty 111

XII The Certainty of Certain Utterances 111

XIII A Perceiving Machine 117

6 Awareness and Consciousness 128

XIV The Ordinary Words 128

XV Awareness and Control 137

XVI Consciousness 142

7 Mental Imagery 149

XVII The Nature of Images and the Introspective Trap 149

XVIII Colours 158

8 Thinking and Reasoning 164

XIX People and Processes 164

XX Reasons and Causes 175

9 Actions and Intentions 184

XXI Intentional Actions 184

XXII Willing 192

XXIII The Importance of Intentional Actions 197

10 Language and Understanding 201

XXIV Knowing and Understanding 201

XXV Language and Information 209

XXVI Conclusions 213

NOTES 217

INDEX 233

PREFACE TO THE ROUTLEDGE

CLASSICS EDITION

It has been a quarter century since I wrote the foreword to the first paperback edition. The cognitive science movement was well underway in 1984 and it is now going stronger than ever, with more and more philosophers playing the sorts of roles in it that I applauded then. Perhaps not so surprising is that in the interim something of a backlash has set in among a small coterie of philosophers who practise resolutely old-fashioned aprioristic philosophy of mind. I have yet to encounter anything worth more than a moment's reflection in that literature, but I grant that it is *logically* possible that something will emerge from it that other researchers on the mind will find worth their attention. I am not holding my breath.

Rereading my book, I have learned something curious about how my own thinking has changed. There are some passages that I find I no longer wish to endorse strongly, but at the same time don't wish to recant. I have moved from conviction to bland agnosticism, and find that I don't even have much to say about why my allegiance has waned. For this very reason, there would be little or no point in my listing these passages; it would be of biographical interest at best and I don't wish to handicap them. So *caveat lector*; some of these passages may be best ignored or they may well be on to something important, may, for all I know, be gems I can no longer appreciate.

There is one set of ideas, however, mainly concentrated in Chapter 4, 'The Ascription of Content' that I do particularly want to commend to the reader. Here I set out for the first time (I think) the ideas that later became known as 'teleofunctionalism,' and drew out the implication that attributions of content were inextricably bound up with evaluations of aptness or appropriateness or rationality, an implication that later was dubbed and deplored as 'meaning holism' by Jerry Fodor. In spite of the

fact that there has been a vigorous campaign by Fodor and others against these ideas, they continue to be reinvented. Recently, in *LOT* 2 *: The Language of Thought Revisited* (Oxford University Press, 2008), Fodor has

broadened his target and rechristened it *pragmatism*. If you think that

'abilities are prior to theories' or that ' *knowing how* is the paradigm cognitive state and it is prior to *knowing that* in the order of intentional explanation' (2008, p. 10), then you are a pragmatist, according to Fodor's expansive definition, along with not just Dewey and Quine, but

Wittgenstein, Ryle, Sellars, Putnam, Rorty, Dummett, Brandom,

McDowell, Dreyfus, Vygotsky, Piaget, Bruner, and Gibson. Pretty good

company, say I. All these folks have been utterly misled, Fodor insists, by their pragmatism, 'perhaps the worst idea that philosophy ever had' (p. 9).

Doesn't that make you want to learn more? What sweeter encomium for

my favourite idea could there be? Fittingly, the opposite of a pragmatist in this taxonomy is a Cartesian, a label Fodor now proudly sports. So here we have a fine confrontation. Which path should you take, the pragmatist or the Cartesian? I guess I should have belaboured my case for pragmatism

more vigorously over the years, since there manifestly remain pockets of unpersuaded philosophers. But the reason I didn't think I had to do this was that I thought I had sufficiently made the case in this book. I still think so.

Daniel Dennett

Cold Spring Harbor

June 1, 2009

PREFACE TO SECOND EDITION

It is now just twenty years since the first draft of this book was submitted (as a D. Phil. thesis at Oxford), and sixteen years since it was first

published. In the intervening period the field of philosophy of mind has grown and changed enormously, a development that is perhaps made easier to see and appreciate by a reconsideration of the way the problems looked (to me) in the late 1960s.

When I was working on the book, its resolute naturalism and earnest

concern with what science could tell us about the mind struck me as quite pioneering – or quite eccentric, depending on my mood. Philosophers of

mind made something of a fetish of their distance from any empirical

investigations, except of the most informal linguistic sort. Times have changed. Now we have cognitive science. There are now more than a few

philosophers of mind who are vastly more knowledgeable about the brain

than I was then (or am now). A fairly professional knowledge of the other cognitive sciences – psychology, artificial intelligence, linguistics – is now considered a virtual qualification for professional status in the discipline.

So what strikes me now about my book is not its pioneering stand, for

we are almost all naturalists today, but its intermittent naiveté. This is mildly embarrassing, but not nearly so embarrassing as would be the

discovery that I hadn't managed to achieve any advance in outlook over

the years. There are also the unalloyed errors, of course, and these are indeed embarrassing. In fact, the only alterations to the text I have made, save for some typographical errors, are the elimination of a few

preposterous howlers. (A good measure of what has been changed is the

correction in example (4) in the second chapter: it was Ponce de Leon, not Hernando de Soto, who searched for the Fountain of Youth!) The more

substantive errors, some of which I have still not recognized or recanted, alas, are left intact.

The task I set myself in the original Preface was ‘to determine the

constraints within which any satisfactory theory [of the mind] must

evolve’, and judging by the subsequent short span of theory evolution, I give myself high marks, at least for identifying the crucial issues and often even getting the matter right. For instance a voluminous debate on the

identity theory has come – and gone – in the intervening years, leaving us with a residue of a few ‘token-token’ identities and a good deal of

‘eliminative materialism’, with the slack taken up by various sophisticated versions of ‘supervenience’ and accounts of psychology as an irreducible special science which is nonetheless properly deemed materialistic. In

short, we are left with just about exactly the position I maintained (plus some useful sophistications) in Chapter 1.

Chapter 2, ‘Intentionality’, draws heavily on Chisholm and others, and introduces the problem of intentionality that has dominated much recent theorizing. There is little I would change in it today (except for those silly mistakes in the examples, which I have changed). The term ‘intentional

system’ appears several times in the chapter and later in the book, but not with the precise sense I later developed for this term (in ‘Intentional Systems’, *Journal of Philosophy*, 1971, and a number of other papers, all cited in the references of my *Elbow Room*, 1984). How fares ‘centralism’, the recommended theoretical approach to the problem of intentionality

that consists in making an initial characterization of the phenomena to be studied in intentional terms, ‘describing the events to be related in law-like ways using either ordinary, or semi-ordinary, or even entirely artificial Intentional expressions’? Except that no one calls his theory ‘centralism’, it fares well indeed, as the recent discussions of ‘folk psychology’ and its semi-ordinary and artificial alternatives in cognitive science attest. The debates on the ground rules have not diminished, with much attention

being devoted to Putnam’s and Fodor’s methodological solipsism, its

strengths, weaknesses, and rivals. This is one of the areas in which I have been provoked to embellish, adjust, revise and extend my thinking considerably – but not recant. In particular, my claims throughout the book about the relationships between inner and outer, function and meaning, rationality and meaning, and rationality and belief have been supported and wonderfully extended by a number of recent books, especially Ruth Millikan's *Language, Thought, and other Biological Categories* and Robert Stalnaker's *Inquiry* (both MIT Press, 1984).

Chapter 3, on 'Evolution in the Brain', stands up well, I think, in spite of its technical naiveté. The very recent upsurge in enthusiasm among neuroscientists for theories of learning as intracerebral evolution is particularly gratifying. While Edelman at Rockefeller, Changeux in Paris, the 'New Connectionists' in artificial intelligence and others are now developing 'evolutionary' models at a level of empirical detail and sophistication I could not imagine in the 1960s, I am pleased to see that their accounts appeal heavily to the concerns I outlined in this chapter. But it is also true that having said what I said, I simply didn't know what to do next with the ideas, so that the recent developments have opened new horizons for me.

The account of consciousness in Part II has some fairly dramatic shortcomings, in my eyes. The account of introspective certainty has some important and salvageable points (in particular about the identity conditions of states and their relations to reports about them), but also some large confusions, which I have tried to correct in more recent

writings. The distinction I draw in Chapter 6 between two different senses of awareness has been dropped from my later work on consciousness, not

out of a conviction that it was entirely mistaken, but for strategic reasons: formulating it properly did not promise to be worth the time and effort.

Recent discussions, however, have convinced me that *something like* that distinction is indeed a strategic necessity if various confusions are to be avoided, so I plan to refurbish a version of the awareness distinction in forthcoming work. Chapter 7, on mental imagery, has been almost entirely swept aside by subsequent empirical and theoretical work on the topic, but perhaps it is useful as an extremely simple and provocative introduction to the issues that are currently being explored. The last three chapters, on thinking and reasoning, action and intention, and language and

understanding provide some foretastes of more recent discussions, and

seem to me today to be not obviously wrong anywhere, but perhaps only

because they are less detailed and ambitious than much current work on

these central topics.

My own views of personal identity over time, and of responsibility,

permit me to take a rather distanced and objective view of this book and its callow author. I find that all things considered I am glad it was written as it was when it was, and also glad that it is now being made available again, this time in a paperback edition. I learned quite a lot from rereading it, and hope that others will find it informative as well.

Daniel C. Dennett

Tufts University

March, 1985

PREFACE TO FIRST EDITION

Books attempting to tell the whole story of the mind have become rarer in recent years, for good reason. No one can hope to master the details, both of empirical data and of theoretical or conceptual nuance. In the face of staggering complexity, prudence has dictated to the student of mind that he must specialize – in the physiology of the nervous system, or in

mathematical models of learning processes, or in the logic of key concepts such as belief, attention, or pain. This retreat from generality has been productive, but has left certain fundamental and pressing questions

virtually untouched. What is the relation between a man's mental life and the events in his brain? How are our commonplace observations about

thinking, believing, seeing, feeling pain to be mapped on to the discoveries of cybernetics or neurophysiology? These questions are important; their answers promise to bridge the specialties and consolidate their gains. But if attempts to answer them are confined, as they largely have been in the past, to philosophical guesswork on the one hand and the speculative

perorations of retiring professors of neurology on the other, no adequate answers will be forthcoming and there will be no unification of theory.

In examining these broad questions of mind and body, I do not try in

this book to tell the whole story, but to set out the conceptual background against which the whole story must be told, to determine the constraints within which any satisfactory theory must evolve. The book specializes by slicing a cross-section, as it were, at a ninety-degree angle to the other specialties. Limiting the task in this way does not rule out all the risks of generality, however. Ideally, anyone hoping to work effectively in this area would have to keep abreast of half a dozen different scientific fields in addition to the advances in the philosophy of mind, but this is out of the question, so I have leaned heavily on résumés of research written for the non-specialist, scientists' gossip about current work, and especially the patient guidance of several colleagues in the different specialties. I have tried to couch all discussion of scientific matters in layman's terms –

indeed it is only in layman's terms that I can understand it myself – and this course has side benefits as well as shortcomings. On the debit side, by the time any bit of science can be rendered in layman's terms it is usually

a bit out of date, which, added to the time-lag of publication, isolates the discussion from the true frontiers of research. This has its bright side, however, for we should not want our working framework for theory to

stand or fall on the often evanescent results hot off the presses of the learned journals.

Part I concentrates on the most general constraints governing scientific theories of the mind, and develops the notion of a distinct mode of

discourse, the language of the mind, which we ordinarily use to describe and explain our mental experiences, and which can be related only

indirectly to the mode of discourse in which science is formulated. In

Chapter 1, a position of ontological neutrality is developed, which allows us temporarily to suspend decision of what ultimate ontological or

metaphysical shape our theory must take, materialistic, dualistic,

interactionistic, vitalistic, etc. This allows certain sterile philosophical conundrums to be avoided, but leads directly to the most powerful

challenge to unification in the theory of mind, the Intentionalist thesis that the mental mode of discourse is ultimately incompatible with the physical mode, and that no translations, reductions or unifications are logically possible. This challenge is examined in Chapter 2, and it is concluded that the best hope for unification lies with the development of a 'centralist'

theory of mind. A centralist theory, in contrast to a peripheralist theory, would attempt to explain and predict human behaviour and experience by

invoking central, internal states and conditions as crucial intervening variables in an explanation couched not in terms of mere stimulus and

response, but in terms of purposive, conscious action. Chapter 3

determines some of the conditions of success for centralism and sketches a theory designed to meet these conditions. The essential task of centralism is seen to be justifying an interpretation of a physical system as a system whose states or events have meaning or content, and in Chapter 4 the conditions for such a justification are examined in detail, and the theory sketch is elaborated to meet these conditions. This leads to a general view of the relationship between the physical, mechanistic side of the story of the mind, and the non-mechanistic account embodied in our ordinary

discourse about people.

In Part II, the bridge built in Part I is exploited in an analysis of consciousness, the feature of mind that is most resistant to absorption into the mechanistic picture of science. Chapter 5 gives a new account, based

on the results of Part I, of the certainty of our introspective access to the

'arena of consciousness', and Chapter 6 analyses consciousness into several separable phenomena. Our ordinary view of consciousness is seen to be muddied by several sets of connotations that deserve separate

treatment, and in Chapters 7, 8 and 9 these are given the attention they deserve. Chapter 10 shows that certain unavoidable imprecisions in the formulation of centralism are inherent in that part of our given conceptual scheme that deals with people and their minds. There is a recurring theme running through the book that traditional analyses, both philosophic and scientific, have failed by postulating unanalysed elements having the very capacities to be analysed, thus postponing true analysis.

My considerable debts to a small number of writers will be evident

from the frequency with which their names appear in footnotes. Others

have helped more directly by reading drafts and making suggestions. First, my thanks and admiration go to Gilbert Ryle, under whose tolerant

supervision the ideas for this book first took shape, and whose always insightful comments led me gently back from many false starts. Then to B. A. Farrell, Nicholas MacIntosh and J. Z. Young, who provided early guidance into the literatures of psychology and neurophysiology, and A. J. Ayer, Dennis Stampe and Jeffrey Sicha, who forced a number of my philosophical ideas into clarity. More recently, my colleagues and students at Irvine have provided valuable assistance, in particular, Gordon Brittan, Karel Lambert, James McGaugh, Julian Feldman, and Frank McGuinness.

Ted Honderich's constructive criticisms of the penultimate draft led to many important revisions. I am indebted also to Henriette Underwood, Eva McCusker and Ida Brown for typing and excellent editorial suggestions beyond the call of duty. Part of my work on this book was supported by a grant from the University of California Humanities Institute, for which I wish to express my gratitude. I also wish to thank the editors of *Behavioral Science*, *Journal of Philosophy* and *Philosophy and Phenomenological Research* for permission to reprint with alterations parts of my articles published by them.

D.C.D.

Irvine, 1968

Part I

The Language of Mind

1

THE ONTOLOGICAL PROBLEM

OF MIND

I THE MIND AND SCIENCE

Those who are convinced of the futility of philosophy are fond of pointing to its history and claiming that there is no progress to be discerned there.

In no area of philosophy is this claim easier to support than in philosophy of mind, the history of which, when viewed through a wide-angle lens,

appears to be a fruitless pendulum swing from Descartes' dualism to

Hobbes' materialism, to Berkeley's idealism, and then back to dualism,

idealism and materialism, with a few ingenious but implausible

adjustments and changes of terminology. The innovations of one

generation have been rescinded by the next so that despite a growing

intricacy of argument and a burgeoning vocabulary of abstruse jargon,

supplemented in each era by the fashionable scientific terms of the day, there have been no real and permanent gains.

The question that defined the pendulum is what the relation is between

mind and body, and the problem that set the pendulum in motion was

Descartes' dilemma of interaction. If, as seems plausible at first glance, there are minds and mental events on the one hand and bodies and physical events on the other, then these two spheres either interact or not. The initially reasonable suggestion that they do interact leads, however, to an impasse of such difficulty that it can be held to be the *reductio ad absurdum* of dualism, at least of the Cartesian variety. If, *ex hypothesi*, mental events are non-physical, they can involve no physical energy or

mass, and hence cannot in any way bring about changes in the physical

world, unless we are to abandon the utterly central principle of

conservation of energy and all its ramifications. Something must give way in this dilemma, and there are many choices available, all traced out by the swings of the pendulum. One can abandon the principle of conservation of energy, and this gives rise to the family of views of non-physical causes and 'occasions'; or one can preserve the principle and deny one of the

other steps that lead to the dilemma. That is, one can deny that there are bodies and physical events and be an idealist, or deny that there are non-physical minds and mental events and be a materialist or physicalist, or hold for a dualism without interaction, and be a parallelist or

epiphenomenalist.

The deficiencies of each of these alternatives, in each of their variations, have been well demonstrated time and again, but this failure of philosophers to find a satisfactory resting spot for the pendulum had few if any implications outside philosophy until recent years, when the

developments in science, especially in biology and psychology, brought

the philosophical question closer to scientific questions – or, more

precisely, brought scientists closer to *needing* answers to the questions that had heretofore been the isolated and exclusive province of philosophy.

Although one can still find in the current literature of the neurologists the old disclaimers about 'leaving to the philosophers' the 'mysteries' of

consciousness, the 'initiation by the mind of neural activities' and so forth, these efforts to skirt the difficult questions are no longer

satisfactory. We need answers now not only to the 'strictly philosophical'

conceptual questions of mind, but to the still quite abstract questions that bridge the gap between physiological theory and the philosophical

understanding of mental concepts.

This gradual and hard-won approach of science to the philosophical questions of the mind-body problem has led to a reshaping by some philosophers of the central concern of the philosophy of mind. In deference to the development of science in the area, they take the task to be providing a satisfactory status for minds and mental events relative to the scientific corpus, and quite naturally their favoured solution to the problem is the identification of mental entities with physical entities. 1 The motive for this identification can be roughly characterized as falling in the same class as the motive for identifying flying saucers with swamp gas, or mermaids with manatees: to avoid ontic bulge. To suppose that there are flying saucers or mermaids in addition to the more ordinary things we hold to exist is to force an inelegant and inexplicable bulge in the shape of our scientific image of the universe, and would eventually force an entirely unwanted revision in some fundamental and otherwise acceptable laws and principles of the natural sciences. Similarly, these philosophers have feared that the assumption of explicitly non-physical mental things – such as thoughts, minds, and sensations – jeopardizes in a more serious way the integrity and universality of the going scientific scheme. Putting their faith in the going scheme, they have determined to identify mental things with, or 'reduce' them to, physical things.

They see as the only alternatives either an asymmetrical scientific picture which includes, in one small corner of the universe, basically *different*, non-physical entities which do not fall under the laws of physics and thus force either drastic changes in these laws or an unsatisfying abridgment of their universality; or the prospect of discovering or proving these mental entities and events to be nothing more than some as yet undescribed physical entities and events, presumably in the brain. If they are right in supposing these to be the only available alternatives, then the

attempt to reduce away the offending mental things is certainly the more reasonable first avenue, just as one should turn to hypotheses about

spacemen, fifth dimensions and anti-gravity machines only after all

attempts to identify flying saucers with more mundane entities have

failed. Unfortunately, the plausibility of the identity theory derives almost entirely from the implausibility of its alternative; if the dangers in denying it did not seem so patent, few would be inclined to suppose that a thought or a pain or a desire just *was* a brain process. In this respect modern identity theory looks all too much like its materialistic predecessors on the pendulum swing: a meta-physically extravagant and implausible monism

into which one is driven by the recognition of the dilemmas in equally

extravagant and implausible dualisms. The identity theory, I shall argue, is wrong, but this does not force us back on to any of the old dualisms, which are equally hopeless. The way out of this unpromising situation is to get off the pendulum entirely, and this involves showing that one of our initial assumptions is not so obvious as it first appears, viz., the assumption that there are minds and mental events on the one hand and bodies and physical events on the other.

II EXISTENCE AND IDENTITY

The strategy that promises to break the spell of the old isms was first exploited by Ryle in *The Concept of Mind*. 2 Ryle argued that mind and matter were in different logical ‘categories’, and *since* they were in different categories there was something logically or conceptually otiose in attempts to identify mind with matter, or in worrying when these

attempts failed, as they must. This line is attractive if it can be made to work: it excuses the identity theorist from his dubious task and tells us at the same time that fear of an ontic bulge is misplaced in this instance. The conceptual elbow room it provides, however, must not be taken to

establish the plausibility of its premises. Is there in fact any logical or conceptual distinction between mental entity terms and physical entity

terms that could be used to justify the claim that the identity theory is a

‘category mistake’?

Illustrations of the sort of differences needed to sustain this sort of claim are not hard to find. Common nouns in English exhibit marked

differences in the ranges of verbal contexts in which they can properly, significantly appear. For example, although ‘sit on the table’, ‘sell a table’,

‘covet that table’ and ‘cut the table in half ’ are all unexceptional, there is something wrong with ‘sit on the opportunity’, ‘sell a twinkle in the eye’,

‘covet the cube root of seven’ and ‘cut the acquittal in half ’. 3 By and large, words for everyday middle-sized objects fit in the greatest variety of contexts while ‘abstract’ and ‘theoretical’ words are the most restricted.

Probably our native ontological bias in favour of the concrete over the abstract derives from this difference in contextual scope; the more

contexts a noun is at home in, the more real, thinglike, and familiar the entity seems. (It should be possible to confirm this bit of speculation about our intuitions and preferences, but confirmation would be irrelevant to our undertaking; no substantive ontological questions could be settled by

appeal to public opinion polls.) There are a few extreme cases in English of nouns restricted to a mere handful of contexts, or even just one. Quine mentions ‘sake’ in ‘for the sake of ’ and ‘behalf ’ in ‘on my behalf ’. 4 Other such idioms are ‘by dint of ’ and ‘plight one’s troth’. You cannot do

anything with or against anyone’s sake, nor can you hope for a behalf,

avoid a dint, or watch over one’s troth. As Quine points out, these degenerate nouns have no combinatory function on their own but are

locked into their idioms. The whole idiom functions as one word, and

there are really only etymological and aesthetic reasons for dividing the idioms typographically at all. This means that any logical or semantical analysis of ‘for my sake’ or ‘on my behalf ’ based on the similarities these share with ‘for my wife’ and ‘on my head’ would be an error bred of

unfamiliarity with the language. Anyone foolish enough to search a house in an effort to find its owner’s sake, or to attempt to identify a man’s behalf with his body temperature or bank balance would be making an

error akin to that of the man who expects a van in each caravan, wonders where the ward is when one marches forward, or expects an audible clank when the dying man finally kicks the bucket. In these cases the Rylean

argument is obvious: e.g., it would be a category mistake for the

physiologist to try to isolate and identify the *dint* of certain muscular exertions, which does not mean that a dint is a secret, non-physical

accompaniment to those exertions.

Similar arguments can be made for less extreme cases. Quine suggests

that nouns for units of measure, such as ‘mile’ and ‘degree Fahrenheit’, are best viewed as integral parts of a small group of idioms rather than as full-fledged nouns which pick out distinguishable items in the world. 5

Once the modern materialist recognizes the limits English places on

‘mile’, he will not be disturbed by the realization that the miles between the earth and the moon are not to be identified with any intervening rays, atoms or trails of plasma. 6

If these trivial cases of category mistakes are well established, their very triviality may seem to weigh against the plausibility of any analogy that would ally them to the rich and versatile vocabulary of the mind.

Thoughts and pains and desires seem to have a much more robust

existence than sakes and miles; although one cannot see or spill ink on a thought or a pain or a desire, a thought – like an explosion – can happen; a pain – like a flame – can be intense; and a desire – like a piece of garlic –

can cause an upset stomach. If the analogy between such terms and our

trivial examples is strong enough both to bar and excuse them from the

crucial identity contexts, this is far from obvious. Certainly it has not been obvious to the many writers who have attempted to defend versions of the identity theory in the last decade.

What is not obvious may nevertheless be shown in the end to be justified, or at least worth investigating. Consider one more example,

closer to the problems of mind in its complexity but without the burden of ancient mysteries. We say ‘I hear a voice’, ‘he has a tenor voice’, ‘you’ll strain your voice’ and ‘I have lost my voice’. Now is a voice a thing? If so, just what thing is a voice? The voice we strain may seem to be as

unproblematic a physical part of the body as the back or eyes we strain, perhaps the vocal cords; but surely one does not have tenor vocal cords or enjoy Sutherland’s vocal cords, or lose one’s vocal cords, and one’s voice, unlike one’s vocal cords, can be sent by radio across the seas and survive one’s death on magnetic tape. Nor does one strain or recognize or lose any vibrations in the air or manifold of frequencies. It might be argued that

‘voice’ is ambiguous – perhaps with some neat and finite list of meanings, so that the voice that changes or is strained is a part of the body, and the voice one enjoys or recognizes or records is some complex of vibrations.

Then what is the voice one loses? A disposition, perhaps. Dividing the

word into these different senses, however, leads us into ludicrous

positions: Sutherland’s voice on the record is not (numerically) the same voice as the one she strained last month, and the voice that is temporarily lost is not the voice we recognize. How many voices does Sutherland

have? If we took this claim of ambiguity seriously, the sentence

‘Sutherland’s voice is so strong; listen to the purity of it in this recording of it that I made before she lost it’ would be a grammatical horror, with each ‘it’ in need of a different missing antecedent, but there is obviously nothing wrong with the sentence aside from a bit of repetitiveness. When the word is viewed (correctly) as unambiguous, attempts to delineate any portion or portions of the physical world which make up a voice will be fruitless – but also pointless. A voice is not an organ, disposition, process, event, capacity or – as one dictionary has it – a ‘sound uttered by the mouth’. The word ‘voice’ as it is discovered in its own peculiar

environment of contexts, does not fit neatly the physical, non-physical dichotomy that so upsets the identity theorist, but it is not for that reason a vague or ambiguous or otherwise unsatisfactory word. This state of affairs should not lead anyone to become a Cartesian dualist with respect to

voices; let us try not to invent a voice-throat problem to go along with the mind-body problem. Nor should anyone set himself the task of being an

identity theorist with respect to voices. No plausible materialism or

physicalism would demand it. It will be enough if all the things we say about voices can be paraphrased into, explained by, or otherwise related to statements about only physical things. So long as such an explanation

leaves no distinction or phenomenon unaccounted for, physicalism with

regard to voices can be preserved – without identification of voices with physical things.

Before trying to fit the vocabulary of the mind to the model of ‘voice’, we should examine our model more closely. Of particular importance is

the question of what ontological distinctions to associate with the

distinctions of verbal function we have examined. In short, are there

voices? One is inclined to answer ‘Of course! We hear and enjoy and recall and recognize voices, so there are voices’, but why the conclusion? Are there sakes? I can do something for Sam’s sake, and he can want

something for the sake of the nation, but then must there be sakes? There seems to be point, and truth, in saying that there really are no sakes or dints, rather less point and truth in denying the existence of miles and degrees Fahrenheit, and a great deal that is implausible in denying the existence of voices. Where then should we draw the line?

It may seem that the line is naturally drawn by determining whether the existence contexts, ‘there are …’, ‘there was …’, and so forth, are

legitimate contexts for the noun in question. This course would establish that there are no sakes (or better: ‘sake’ does not denote or name or refer to anything; if ‘there are sakes’ is to be a solecism, its negation must be as well). ‘Mile’ would also be ruled to have no ontic force, since, for

example, ‘there are seven miles between …’, ‘there once was a mile …’

and ‘there is a mile …’ are all improper. Voices would be admitted,

however, on the strength of sentences like ‘there was a voice in the dark and I recognized it’. 7 The justification for this course would be slender enough even if our grammatical intuitions in particular cases were strong and unanimous, but they are not. Our intuitions are poor witnesses just when they would be most heavily relied on. Consider the claims:

(1) ‘there is a mile between them’ is deviant usage

(2) ‘there is a mile to go’ is not deviant usage

(3) ‘there are five miles of hard hiking between the peaks’ counts as

affirming the existence of hiking, not miles.

Are these any easier to assess than the question they would be expected to settle, viz., should we say, in the context of discussing or choosing

ontologies: there are miles? As these claims get harder to assess, their value as criteria wanes on two fronts: relevance and decidability.

A more lenient ontology could be loosely put: something for every noun

(and noun phrase, etc.). This most relaxed course cheerfully admits the existence of dints and sakes, and treats the whole question of ontology with what some may hold to be deserved disrespect. But consider the

following exchange:

‘How old is Smith’s sake?’

‘Sakes don’t exist in time.’

‘But they do exist, don’t they?’

‘Why not?’

‘Then if Smith’s sake is timeless, we’ll be able to do things for

it after he’s dead.’

‘No; although a sake is timeless, it can no longer receive

benefits after the death of its owner.’

‘Then I might only think I was doing something for Smith’s

sake, if all along he was dead without my knowing it?’

‘No, you’d be doing it *for* Smith’s sake, only his sake would no longer have any *use* for whatever you were doing.’

This sort of nonsense should be blocked one way or another. If one merely forbids the word to appear in the various *syntactical* roles it appears in above, what is one clinging to when one refuses to admit that the word has no *ontological* role? Asserting existence under these conditions is as empty

as denying it. Answering yes or no to an ontological question only begins to have some point when we have decided that granting the

existence of something licenses us to ask (and expect answers to) certain very general questions about it, e.g., what sort of thing is it?, does it exist in time?, and especially, is it identical with x? This last question is indispensable. It is hard to see what anyone could have in mind by

affirming the existence of something, if he then disallowed this question.

For, let us divide our universe into as many different ontological

categories or types as we wish, if we assert that x is a thing existing in sense A, or in category A, and y is a thing existing in sense B, or in

category B, then at the very least we must acknowledge that we have just spoken of *two* things, x and y, not just one – or in other words that x is not identical with y, but is *another* thing.

Consider voices again. We entertained the proposal to admit voices into our ontology because under some circumstances ‘there is a voice …’ rings true in the ear, but there are better reasons for denying them. If the

anatomist or physiologist or acoustician were to be concerned because

among all the things encompassed by his theories there still were no

voices; if he were to suppose this meant he had left something out, something perhaps even inaccessible to science, he would have been

confused by our admitting voices in our ontology. He assumed this meant he could safely reason: Is the voice identical with the larynx? No. Then is it the lungs? No. Is it a stream of air? No. Is it a sound? No. Then it must be some *other* thing I have not yet examined. We must rule out this series of questions, but if we must, it cannot be on the grounds that voices are logically (or ‘by meaning’) non-identical with physical entities, for we cannot rule out a question simply because its answer is ‘No (as a matter of logic)’. We can rule out the questions only by declaring them ill-formed,

and hence admitting no answer. 8 So the ontological question is of a piece with the question whether the odd sentences (e.g., ‘I can sit on an

opportunity’, ‘The voice is identical with the larynx’) are logically false or ill-formed, for although whenever two things exist they may well be

logically (as opposed to contingently) non-identical, so long as we do hold there to be two things we cannot burke the question of identity by

declaring it ill-formed. The point, then, in denying the existence of voices is to permit the claim that physicalists need not identify voices with any physical thing (talk of such identities being ill-formed). That such a denial is to some extent counterintuitive is not contested, but then it is also counterintuitive to suppose there are dints and sakes. Since no drawing of the line is clearly superior in intuitiveness, we may turn to other criteria.

The denial of voices has at least the systematic advantage of providing a reason for ruling out the physiologist’s questions, which are, intuitively, wrongheaded. 9

Certainly no one interested in voices ever fell into the misunderstanding just described, but it is tempting to suppose that not only philosophers but also psychologists, neurophysiologists and cyberneticians are bedevilled at times by a parallel confusion over the ontological status of the mental

vocabulary. The outcome of our analysis of ‘voice’ was the adoption of a relatively restrictive sense of ‘exists’, and this allows a clarifying

reformulation of the mind-body problem: when the neurophysiologist – or his armchair counterpart, the physicalistic philosopher – asks whether he has left *anything* out of his theory of the mind, or whether *anything* relevant to the operations he is studying is outside the domain of his

science, he is asking whether there *exist* (in this strong, restrictive sense) any such things. On the one hand the answer that such things do *not* exist may come as a relief to the neurophysiologist or physicalist, but may also come as a surprise, for even in this strong sense of ‘exist’ it *seems* that pains exist as surely as pins, desires and ideas as surely as electrons. On the other

hand the answer that such things *do* exist will mean that the Rylean strategy has led us back to our starting point; we are back on the pendulum and must decide between the old alternatives of interactionism, parallelism, identity theory and so on. No amount of talk of categories and category mistakes will keep us from the snares of dualism unless we are prepared to grant ontological priority to one category at the expense of another.

Determining the ontological status of the mental vocabulary will not be simple. Instead of a general argument there must be detailed

investigations of individual words and families of words, and we must be alert to the possibility that only a partial case can be made. To expedite this investigation – which will be concentrated in Part I, but will cast lines through all of Part II as well – I wish to introduce a technical term. I shall call nouns or nominalizations that do denote or name or refer to existing things (in the strong sense developed above) *referential*, and other nouns and nominalizations, such as ‘sake’, ‘mile’ and ‘voice’, *non-referential*. 10

Non-referential words and phrases are then those which are highly

dependent on certain restricted contexts, in particular cannot appear

properly in identity contexts and concomitantly have no ontic force or

significance. That is, their occurrence embedded in an asserted sentence never commits the asserter to the existence of any entities presumed

denoted or named or referred to by the term. Our prospects can now be

outlined with the use of this new term. We may find that no mental entity terms can be plausibly claimed to be non-referential, in which case we are thrown back on the old alternatives: either minds are identical with

physical entities or they are not, in which case we must put together some

∴

∴

sort of dualism. Or we may find that the entire vocabulary of the mind

succumbs to non-referentiality; this would allay all our fears of ontic bulge and leave the neurophysiologist in the same relatively

uncomplicated position as our voice-investigator. Or we may find that the mental vocabulary is a mixed bag; in this event the crucial question will become whether or not the referential terms in the mental vocabulary refer to things identical or non-identical with physical things. In this way we might be able to put together a theory that was throughout physicalistic in import, but only an identity theory with respect to some of the mental

terms: viz., the terms that refer to things that actually exist.

Once we decide that a term is best viewed as non-referential, we *fuse* it in its proper contexts, as we noted earlier with ‘sake’ in the irreducible idiom ‘for-the-sake-of ’. The contexts maintain their significance but are not subject to further logical analysis; their parts become like the ‘table’

in ‘potable’. The chief advantage of fusion is the ontological absolution we gain, but there is a price to pay. For example, were we to take the voice problem seriously and proceed with a zealous and rigorous analysis of

‘voice’-idioms, we should have to accept that ‘John strained his voice’ is not to be treated as an instantiation of the formula ‘x strained y’, for now

‘strained-his-voice’ is a fused context not open to further analysis. This means we shall no longer have any logical licence for the apparently sound inference:

Anyone who strained anything was doing something excessive

John strained his voice

John was doing something excessive.

But we must be willing to pay this price if we are to deny a licence to the inference:

The only thing John strained was his vocal cords

John strained his voice

John’s voice is identical with his vocal cords.

Even more awkwardly, fusion will often extend beyond a few words to left and right of our non-referential term. For example, fusion must extend to

any pronominal cross-references to voices. Most implausibly, for example, the whole sentence

‘The first thing about his *voice* that struck me was that I had heard *it* before’

becomes impenetrable to logical or semantical analysis. This conclusion will seem preposterous until we reflect on just what is and is not being prohibited by fusion. Obviously fusion does not prohibit analysis the way a dictator prohibits free assembly; it merely forbids certain sorts of

interpretations being put on the results of analysis. It may well be possible to produce a ‘semantics’ for the ‘voice’-idioms, and a ‘logic’ as well; were this accomplished the only thing fusion would prohibit would be any

attempt to treat this ‘semantics’ as an *extension* of the semantics of our referential vocabulary, with voices as a *sort* of thing *in addition to* the sorts of things referred to by referential terms. From the vantage point of our base camp in the midst of existing things and referential terms, ‘voice’

must forever be non-referential; only in this way can the alternative of identity or non-identity be denied. Provided this crucial bit of insulation is maintained, however, there is no limit to the sort or number of systems one may erect for dealing with the parts of fused expressions, and it is even to be expected that any systems discovered will be virtually parallel to the semantics and logic on the referential side of the divide. 11 Fusion, then, is *from a point of view*; it renders contexts impenetrable only from certain angles, as it were. This impenetrability is not just a hindrance. It also provides a degree of freedom by excusing the analyst from finding all our logical and semantical rules obeyed on the far side of the fusion

barrier.

It has not yet been decided that all or any of our mental entity terms *are* non-referential, but only that we should investigate to see. To this end I propose to employ a tactic which can be called *tentative fusion*. We wish to proceed with no ontological presuppositions to the effect that mental entity terms either are or are not referential, and this can be accomplished by treating all sentences containing mental entity terms as tentatively fused, subject to further discoveries which will lead us to confirm the fusion or relax it. We do not *assume* from the start, that is, that certain sorts of questions have answers, that certain sorts of implications hold,

that certain sorts of parities exist between physical entity nouns and mental entity nouns. What we start with, then, are sentences containing the mental entity words to be examined. We may say these sentences are 'in

mental language', and we acknowledge that as wholes they are significant and hence true or false. Part of what is then at issue is whether or not the parts of these sentences should be construed to fall under our standard semantics – whether or not to relax the fusion. 12 The broader question of which this forms a part is whether or not these sentences, accepted either as wholes or as analysed, can be *correlated in an explanatory way* with sentences solely from the referential domain of the physical sciences. Our model here is the case of voices; the explanation of vocal phenomena may contain no *reference* to voices; can the explanation of mental phenomena similarly avoid reference to minds, thoughts, pains? By taking whole

sentences as our initial units we avoid making the one presupposition that leads irresistibly to the pendulum of old-fashioned alternatives: the

presupposition that 'mind', 'thought', 'pain' are referential, or in other words, the presupposition that there are minds and mental events on the one hand, and bodies and physical events on the other.

Not just any mapping of sentences on sentences will constitute an

explanatory correlation, of course. One could associate each true sentence of the mental language with a sentence which catalogued as exhaustively as possible the entire physical state of the person or persons in question, but

this would explain nothing. At the very least the sentences associated with the mental language sentences must describe conditions which vary

in systematic ways related to distinctions in the mental language

sentences. The degree of freedom, however, which we obtain by

tentatively fusing the mental language sentences, will allow us to avoid one preposterous requirement of crude identity theory. As Putnam points out, the supposition that a particular mental experience, e.g., thinking of Spain, is identical with a particular physical state requires that all beings truly said to be thinking of Spain must be in this particular physical state, which rules out, most implausibly, the possibility that beings with a

different biochemistry from ours, or a differently embodied nervous

system, could think of Spain. 13 Even among *homo sapiens* it is not plausible to insist that when two of them are both thinking of Spain they must share some unique physically describable state.

Rather than attempt to characterize in an abstract fashion the minimum requirements of a satisfactory explanation, let us proceed to see what

correlations we can find. Once they have been set out we can ask whether or not they constitute an adequate explanation of mental phenomena. In

most general terms our task is to provide a scientific explanation of the differences and similarities in what is the case in virtue of which different mental language sentences are true and false. Thus, for example, our task is not to identify Tom's thought of Spain with some physical state of his brain, but to pinpoint those conditions that can be relied upon to render the whole sentence 'Tom is thinking of Spain' true or false. This way of

proceeding still characterizes the task as finding an explanation of the mind which is *unified with, consistent with*, indeed a *part of* science as a whole, but eschews – at least initially – the obligation to find among the *things* of science any referents for the terms of the mental vocabulary. This obligation

will only be taken on in the event that some or all of the mental terms resist all efforts to treat them as non-referential.

The first obstacle thrown in the way of our attempt to achieve

explanatory correlations is a very general, but very powerful, argument to the effect that those features of the world in virtue of which certain mental language sentences are true or false are outside the domain of the physical sciences, and not describable or subject to explanation within the scientific framework. If the argument is sound then our having reached an

appropriately noncommittal stance for dealing with the ontological

problem will be to no avail, for this argument is concerned with relations between sentences. This argument will be presented and examined in

Chapter 2.

2

INTENTIONALITY

III THE PROBLEM OF INTENTIONALITY

In the previous chapter we formulated a stance that enables us to ask what the relation is between the physical sciences and the truths expressed in our mental language while carrying the minimum of metaphysical

baggage. We avoid all ontological presuppositions about mental entities by tentatively treating all sentences of the mental language as containing no referential terms. Thus for at least the time being we absolve the scientist from the responsibility of discovering physical events, states or processes which deserve to be called thoughts, ideas, mental images and so forth. No entity on his side of the fence need line up with mental language in such a way that we would say he has discovered what thoughts are, or isolated a mental image or even the experience of having-a-mental-image. We have

the mental language, and since the suggestion that all the things we say in the mental language might be false is incoherent, we also have the *truths*

expressed in mental language. The task is to relate these truths to the scientific corpus, and further to explain the relations. Since we cannot very well claim to have explained a mental phenomenon if we are unable to say (in the scientific language of our explanation) when a sentence heralding the occurrence of the phenomenon is true and when not, our task will

involve at least this much: framing within the scientific language the

criteria – the necessary and sufficient conditions – for the truth of mental language sentences. At this point we face a very general argument

designed to show that this is impossible, that no criteria for mental truths can be expressed in the language of science.

The nineteenth-century psychologist-philosopher Franz Brentano, in his

Psychologie vom Empirischen Standpunkt, claimed to have formulated an exact and useful distinction between mental phenomena and physical

phenomena, and it is this distinction that forms the basis of the argument.

Mental phenomena, according to Brentano, exhibit Intentionality, [1 a](#) term he revived from medieval philosophy. ‘Every mental phenomenon is

characterized by what the scholastics of the Middle Ages called the

Intentional (and also mental) inexistence (*Inexistenz*) of an object (*Gegenstand*), and what we would call, although in not entirely

unambiguous terms, the reference to a content, a direction upon an object

(by which we are not to understand a reality in this case) …’[2 B](#)rentano’s thesis divides roughly into two parts – although how clearly Brentano saw the division is hard to say. Some mental phenomena are ‘directed upon’ an object (and these objects have unusual characteristics), and other mental phenomena are related to a content or proposition or meaning. There is

some difficulty, as we shall see, in welding these two parts into a single characteristic of Intentionality, and yet intuitively Brentano’s insight is

about one characteristic, and an important one.

Subsequent investigators have annexed to this distinction the claim that no statement or statements about non-Intentional phenomena can have the same truth conditions as any statement about Intentional phenomena.

Since, roughly speaking, the domain of statements about Intentional

phenomena is the domain of statements expressed in what I have called

the mental language, and the domain of statements about non-Intentional phenomena includes all discourse in the physical sciences, the force of this claim is that no systematic correlation of sentences of the sort envisaged in the previous chapter is logically possible. As Brentano's

characterization of the distinction between Intentional and non-Intentional is clarified and modified, the strength of this claim will become evident, and unless a way of refuting it is found, the metaphysical brush-clearing of the previous chapter will have been to no avail, since the Intentionalist thesis, quite independently of any arguments about the ontological status of mental entities, proclaims an unbridgeable gulf between the mental and the physical.

Brentano's point about *direction upon an object* is this: One cannot want without wanting something, imagine without imagining something, hope

without hoping for something, and yet the *object* in all these cases does not, or need not, exist in the fashion of objects of physical actions, such as lifting, touching, sitting upon. Thus if I want a wife it not only does not follow that there is a wife I want, but also does not follow that there is a woman whom I want to marry, any more than it would follow from the fact that I want a space ship that there is a space ship that I want. On the other hand, if I hit a wife or take a ride in a space ship, it follows that there is a wife I hit and a space ship in which I ride. The objects of wanting are thus said by Brentano to have Intentional inexistence, and the same holds true for the objects of imagining, remembering, hoping and so forth. I can

imagine a sphinx, remember the dead, and hope for a cure to the common

cold, none of which objects exist – at least in the ordinary way. Brentano says these objects have ‘inexistence’ but it is not altogether clear whether Brentano meant by his prefix ‘in-’ that these objects enjoy a form of *non-*

being, or existence *in* the mind, or both (Cf. Anselm’s ‘ *in intellectu*’) – but in any case it is a queer sort of existence he had in mind.

His point about *relation to content* is that in addition to believing in ghosts (a case of direction upon an object) we also believe *that* …, and hope *that* …, and in these cases there is not so much an object directed upon as a proposition related to. It was, perhaps, but certainly should not have been, Brentano’s thesis that these propositions stand in the same

relation to mental phenomena as the Intentionally inexistent objects

mentioned above, that believed propositions enjoy the same kind of queer existence in relation to mental phenomena as hoped-for rains and

believed-in ghosts. A moment’s reflection on the different status of the believed-in ghost and the believed proposition ‘there are ghosts’ should convince us that this is a blind alley. In fact, Brentano’s proclivity to talk in terms of strange *objects* is in general more trouble than not, and an effort will be made to make his point entirely in terms of relation to

content. 3

The first step in reformulations of the Brentano thesis has usually been to raise the subject level of the discussion from phenomena to talk about phenomena, turning the distinction into a matter of how we describe or

allude to certain phenomena in our ordinary language. Thus Chisholm says that ‘we can formulate a working criterion by means of which we can

distinguish *sentences* that are Intentional, or are used Intentionally, in a certain language from *sentences* that are not’4 (my italics). This procedure coincides nicely, of course, with our deliberate ontological blindness; we do not *suppose* that there *are* any actual phenomena (thoughts, beliefs, desires)

for Intentional sentences to be about. Chisholm proposes three independently operating criteria for Intentional sentences. 5

(1) A simple declarative sentence is Intentional if it uses a substantival expression – a name or a description – in such a way that neither the sentence nor its contradictory implies either that there is or that there isn't anything to which the substantival expression truly applies.

For example, neither 'I want a space ship' nor 'I do not want a space ship' implies either that there is or that there is not a space ship, and hence both sentences are Intentional.

(2) Any noncompound sentence which contains a propositional clause … is Intentional provided that neither the sentence nor its contradictory implies either that the propositional clause is true or that it is false.

For example, neither 'I hope that it will rain' nor 'I do not hope that it will rain' implies that it will or will not rain, and hence both sentences are Intentional.

(3) If A and B are two names or descriptions designating the same thing or things, and sentence P differs from sentence Q only in having A where Q has B, then sentences P and Q are Intentional if the truth of one together with the truth that A and B are co-designative does not imply the truth of the other.

A familiar example is Quine's: although Tully is identical with Cicero

('Tully' and 'Cicero' name the same individual), from 'Tom believes

Cicero denounced Catiline' it does not follow that 'Tom believes Tully

denounced Catiline' is true, since Tom may not know or believe that Tully and Cicero are one.

Chisholm's three criteria come close to reproducing Brentano's

distinction, but a few alterations must be made to meet a host of apparent counterexamples. 6 First, sentences containing such verbs as 'hunt' and

'search', which are not obviously *mental* terms, are usually construed to meet (1), as

(4) Ponce de Leon was searching for the Fountain of Youth shows.

The usual line with these, which I shall follow, is to reconstrue 'mental'

rather more broadly, or replace it with 'psychological', and accept such sentences as falling within Brentano's notion of Intentionality and on a par with the sentences with a more obviously mental subject matter. A point of contact can be seen here between philosophical and psychological policy.

'The rat is *searching* for an escape route' is as much to be avoided by the rigorous behaviouristic psychologist as 'The rat *hopes* or *believes* …', for

'search', in virtue of its Intentionality, is a *non-observational* term, just as much as the more obviously mental terms, and hence has no place in the

'pure' data language of the behaviourists. If a term like 'search' is to be used at all by these psychologists, it must be defined in observational terms, all of which are non-Intentional. The difficulty psychologists have had in providing these definitions comes as no surprise to the

Intentionalist, of course, for he has an argument to show such definitions to be strictly impossible. 7

Counter-examples that force a different adjustment in Brentano's thesis are the case of clearly mental expressions which fail to meet the criteria.

Thus

(5) Tom perceives the cat

implies the existence of a cat, violating (1), and

(6) John knows that Smith is a lawyer

implies that Smith is a lawyer, violating (2). 8 There are other kindred cases, such as 'discover', and 'recognize'; and 'is under the illusion that'

and 'hallucinates', which imply the falsity of the clause or non-existence of the object. 'Fears' might be seen as a fence-sitter, since 'Tom fears the bogyman' might be held to be true without implying the existence of non-existence of any bogyman, while 'Tom fears dogs' or 'Tom is afraid of the dark' might be seen to imply the existence of dogs and the dark. 'Foresee'

in the present tense does not imply the truth of the clause ('She foresees that there will be a great disaster') but 'Noah foresaw that there would be a great flood' might be held to imply the occurrence of the flood.

The idiosyncrasies of these ordinary expressions and their failure to

meet our criteria should neither surprise nor discourage us – nor for that matter should they engage our interest for long. These are clearly mongrel terms, part mental or psychological, part contextual or epistemic. Thus

'know' has in addition to its implications regarding the psychological state or attitude of the knower the implication that what is known is true. And

'John sees the dog' speaks not only of John's state of mind, but also of his

position in the physical environment. We could further muddy the waters by inventing new expressions which would be similar to such paradigm

Intentional expressions as 'believe' and 'want' but violate the criteria.

Thus if ‘fwant’ had the implication that what was fwanted was in one’s

pocket all along, it would violate (1), and if ‘beglieving’ was restricted so that people beglieved that you could square the circle or travel faster than light, it would violate (2).

The best way with these is to allow that they are Intentional in virtue of the fact that they have Intentional implications. ‘John knows that p’

implies ‘John is under the impression that p’, and ‘John perceives x’

implies ‘John seems to himself to perceive x’, and these meet Chisholm’s criteria. 9 This treatment is in some ways parallel to that of Husserl and later Phenomenologists, who speak of the *epoché* or ‘bracketing’ of external, real-world implications, a practice that is held to get us to the true mental phenomena. In addition to the mongrel terms mentioned, there are a host of other expressions with less direct Intentional implications.

Thus ‘John signed the contract’ implies a number of things about John’s beliefs, as do ‘John was introduced to Mary’ and ‘Tom knows Mr. Smith’.

The picture that emerges is one of an arsenal of quasi-mental, quasi-

psychological terms with Intentional components, and these components

can generally be given a satisfactory analysis in terms of the ‘pure’

Intentional terms, such as ‘believes’.

Failure to recognize this situation can lead to spurious philosophical

puzzles. If ‘know’ is held to refer purely and simply to a psychological state, we get the following absurd exchange: ‘I know that John is in

Boston’ ‘But he isn’t. He changed his mind and left Boston yesterday’

‘That’s strange. I could have sworn that I *knew* that, but I see now that I didn’t. *I must have misidentified my mental state*. I must be careful in the future to keep a close watch for the true earmarks of states of knowledge.’

It can be argued that the entire Cartesian epistemology is bedevilled by this mistake. The absurdity of treating the mongrel terms as referring neat to psychological states can be brought out even more clearly in this

exchange: ‘I hate Stalin’ ‘But Stalin is dead’ ‘Oh! Then I guess I don’t hate Stalin, since that would imply his existence.’

Another challenge to our criteria is Geach’s example:

(7) I owe John a horse[10]

the truth of which does not imply the existence or non-existence of a horse, satisfying (1). This can properly be viewed as Intentional, however, in virtue of its psychological implications, concerning promising, stating, believing and so forth. We are not through the counterexamples yet,

however. Quine offers several more:[11]

(8) ‘Tully was a Roman’ is trochaic

does not admit substitution of ‘Cicero’ for ‘Tully’ thus apparently

satisfying (3), and by no stretch of the imagination is (8) about the mental life. (8) also apparently satisfies (2) as

(9) ‘Tully was a German’ is trochaic

indicates, provided we allow the phrase in inverted commas to count as a propositional clause. As Quine points out, however, direct quotation is best viewed as making names of expressions out of expressions, so that the

‘Cicero’-‘Tully’ substitution is an illicit alteration *within* a name, and (8) and (9) are then simply sentences about two different trochaic sentences.

This effectively rules out the family of counterexamples based on direct quotation, but others are generated by the alethic modalities of necessity and possibility.

(10) 9 is necessarily greater than 7

is true, but substitution via the identity '9 = the number of planets' gives us

(11) the number of planets is necessarily greater than 7

which is false. So (10) meets criterion (3), but is not about mental

phenomena. Unlike the others, this counterexample cannot be legislated

away without controversy. On the one hand there are the various systems of modal logic which would construe (10) and its kind in such a way that the implication to (11) would be blocked while still preserving in some form the principle of substitution of codesignative expressions *salva veritate*. [12] On such an analysis necessity statements would no longer meet

(3). [13] Alternatively one can be sceptical of modal logic and take the way out suggested by Quine. [14] One can claim that if 'necessarily' means anything, it qualifies statements or sentences, not, as in the example, relations like *greater than*. (10) then, must be revised to read (12) '9 is greater than 7' is necessarily true

and by our convention for direct quotation the proposed substitution for

'9' is prohibited. If neither of these suggestions can be made to stick, we shall simply have to enlarge the Intentionality thesis to include sentences not about psychological phenomena. This would mean abandoning

Brentano's claim to have discovered the hallmark that distinguishes the mental from the physical, but the crucial argument – that no truth criteria for Intentional sentences can be formulated in the terms of physical

science – will not be abrogated; its conclusion will merely be extended to cover areas outside the mind-body problem. [15]

To sum up, the effect of all these counterexamples is to bend the

Brentano thesis into something quite unlike the original. As adjusted,

Intentionality is not a mark that divides phenomena from phenomena, but sentences from sentences – and whereas Brentano associated Intentionality

with the *mental*, we have given it a broader association with the *psychological*, and are prepared even to abandon this claim if a suitable home cannot be found for the non-psychological modalities. It will help to bend the thesis just a bit more, in an effort to unify Brentano's dual point about objects and content. Chisholm's three criteria preserve this duality: (1) is about direction upon an object, and (2) and (3) are about relation to a content. Criterion (1) could be dropped and the distinction of

Intentionality could be made entirely in terms of the truth and reference of propositional clauses, if we could replace object-sentences with sentences of propositional attitude. Some object-sentences transform quite

gracefully. 'I am hoping for peace' becomes 'I am hoping that there will be peace' and 'I believe in goblins' becomes 'I believe that there are

goblins'. Others are at best awkward. 'I want a wife' might go easily into

'I wish that I had a wife', except wishing and wanting are distinguishable, so 'I want it to be the case that I have a wife' is preferable but

cumbersome. One could as well use 'I want that I should have a wife',

which is reminiscent of Damon Runyan ('You want I should lean on him,

boss?') and so is not entirely non-ordinary; or, even more barbarously, one could say 'I want that I have a wife'. It is not to be expected that our ordinary language will provide natural expressions to serve as translations in every case. Still less is it to be demanded that for every mongrel

Intentional expression there must be an ordinary language expression that satisfactorily analyses the Intentional component of the mongrel

expression, for this would require a sort of systematic perfection in

ordinary language that there is no reason to suppose must exist.

Some object-sentences cannot be *translated* into single propositional-attitude-sentences at all. No propositional paraphrase of 'John hates

spinach', for example, is remotely convincing as a translation. 'John

believes that spinach nauseates him' and 'John wants that he is not served spinach' say both less and more than the object-sentence. 16 A large enough collection of such partially successful paraphrases might serve, either in some strict alternation and conjunction system or in the looser

'family' way, as a suitable transformation of the sentence, but there is no particular payoff in setting out to elaborate these collections for each and every object-expression.

The point of thinking in terms of propositional attitudes even where no neat sentences of propositional attitude can be produced is that Intentional objects, even under the linguistic interpretation given them here, lead almost inexorably to metaphysical excesses, 17 and the characteristic of these objects that accounts for this is one that it can be argued serves precisely to show that Intentional objects are not any kind of objects at all.

This characteristic is the dependence of Intentional objects on particular descriptions. As criterion (3) indicates, to change the description is to change the object. What sort of thing is a different thing under different descriptions? Not any object. Can we not do without the objects altogether and talk just of descriptions? When John wants the stick on the ground, and the stick turns out to be a snake, it would be wrong or meaningless to say that John wanted the description 'the stick on the ground', for we do not *want* descriptions and so in Brentano's terms his mental phenomenon is not *directed upon* the description, but is it not *directed by* the description? It is suggested that Brentano's thesis might be altered to read

'All mental phenomena are directed by (or simply: related to) unique

descriptions or whole propositions which usually, but not always, have

reference to real objects in the world.' Thus Brentano's thesis becomes,

what it is often supposed without argument to be by writers in the field, simply that mental phenomena differ from physical phenomena in having

a content, or relating to meaning, in the sense that their identity as

individual phenomena is a matter of the unique descriptions or

propositions to which they are related.

Raising the subject level of discussion back up from phenomena to talk

about phenomena, from things to sentences, the point is this: Intentional sentences are *intensional* (non-extensional) sentences. 18 Briefly and roughly, the *extension* of a term is the class of all things of which the term is true, or to which the term refers. Thus the extension of 'Presidents between 1961 and 1968' is the class containing the two members, Kennedy and Johnson. The *intension* of a term is, roughly, its meaning. The terms

'Democratic Presidents between 1956 and 1968' and 'Twentieth-century

Presidents whose names begin with adjacent letters of the alphabet' have the same extension as 'Presidents between 1961 and 1968', but clearly all three have very different meanings or intensions. 'Goblin' and 'sphinx'

have the same extension (the null class), but different intensions. The going scheme of logic, the logic that both works and is generally supposed to suffice for all scientific discourse (and, some hold, all *significant* discourse), is extensional. That is, the logic is blind to intensional

distinctions; the intersubstitution of coextensive terms, regardless of their intensions, does not affect the truth value (truth or falsity) of the enclosing sentence. Moreover, the truth value of a complex sentence is always a

function of the truth values of its component sentences. Criteria (2) and (3) indicate that Intentional sentences do not follow the rules of

extensional, truth-functional logic, and hence they are intensional. This expression of the position leads us to the central claim of the

Intentionalists, that Intentional phenomena are absolutely irreducible to physical phenomena. Put in terms of sentences, the claim is that

Intentional sentences cannot be reduced to or paraphrased into extensional sentences about the physical world. The claim goes beyond the obvious

fact that Intentional sentences are intensional, and hence cannot be, as they stand, extensional – to the more remarkable claim that no sentence or

sentences can be found which adequately reproduce the information of an Intentional sentence and still conform to extensional logic. This is to be contrasted with the situation with respect to, for example, Quine's

sentence:

(13) Giorgione was so called because of his size.

This sentence, intensional as it stands since it does not admit the

substitution *salva veritate* of the coextensive term 'Barbarelli', can be replaced by the sentence

(14) Barbarelli was given the name 'Giorgione' because of his size

which is extensional, since the occurrence of 'Barbarelli' in the sentence is replaceable *salva veritate* by 'Giorgione', and the occurrence of

'Giorgione' is within inverted commas, and hence, by standard

convention, refers to the *name* 'Giorgione' and not the man. Since it refers to the name, 'Barbarelli' is not coextensive with it; its coextensive

brethren include 'the maximizing form of "Giorgio" ' and 'the name

which appears on page seven of the art gallery catalogue', and so forth, and these are all substitutable *salva veritate*. That is, making two legitimate substitutions in (14) we get

(15) Giorgione was given the name which is the maximizing form

of 'Giorgio' because of his size

which has the same truth value as (14). The Intentionalist claim is that no extensional sentence – or longer paraphrase – could reproduce the sense of an Intentional sentence in the manner in which (14) reproduces the sense of (13).

This claim has been argued for in a number of different ways. Since the conclusion involves a negative existential claim, viz., that there are no such paraphrases, it can never be absolutely established, but only made extremely compelling. 19 Quine attempts to generate this conviction by taking the case of indirect quotation or *oratio obliqua* ('x says that p'), and arguing that 'for all its tameness in comparison with other idioms of

propositional attitude, and for all its concern with overt speech behaviour,

[it] seems insusceptible to general reduction to behavioural terms; the best we can do is switch to direct quotation, and this adds information.'20 That is,

(16) He says, 'it is raining'

is not a satisfactory paraphrase of

(17) He says that it is raining

for (16) adds information and hence could in many circumstances be false while (17) was true, and vice versa. 21 One is invited to reflect on the impossibility of there being any physical state of affairs that would be in force always and only when someone was saying that it is raining – an act that might be accomplished on one occasion by merely nodding, on

another by shouting 'es regnet', on another by saying 'You bet it is'. If so overt an activity as saying that something is the case is not subject to behavioural, extensional paraphrase, what hope is there for such hidden, private phenomena as believing and imagining? Quine and Chisholm also

present arguments about believing and intending, of which the central

point is that efforts to provide behavioural analyses of these two

phenomena are doomed by a vicious circle of implications. 22 Take, for example, the belief that it is raining. What behaviour would clinch it that A believes it is raining? No matter what is suggested, it will turn out that this is a clincher demonstrating that A believes it is raining *only* if we assume that A has some particular purposes or intentions. A's saying 'It is raining' or answering 'Yes' to the question 'Is it raining?' only counts as evidence on the assumption, *inter alia*, that A intends not to deceive us and 'intends' is an Intentional idiom. A's finding a tree or roof to stand under is no more evidence, for it depends on A's intending to stay dry. If ascription of belief always depends on an assumed ascription of intention, the converse holds as well. A's intention to stay dry is not behaviourally demonstrated by his cowering under the tree except on the assumption that he believes it is raining, that he believes that he would get wet if he did not stay under cover, and so forth. A survey of the other Intentional and

mongrel Intentional idioms shows that the use of any one of them has

implications about beliefs and intentions, so the circle that prevents a behavioural paraphrase of belief and intention sentences infects the whole realm of the Intentional. It is, of course, no argument against this that behavioural data are 'for all practical purposes' completely reliable as clues to Intentional ascriptions, for we are not concerned here with

practical purposes, but with theoretical foundations.

IV TWO BLIND ALLEYS

The Intentionalist thesis of irreducibility is widely accepted, in one form or another, and there are two main reactions to the impasse: behaviourism and Phenomenology. The behaviourist argues that since the Intentional

idioms cannot be made to fit into the going framework of science, they

must be abandoned, and the phenomena they are purported to describe are claimed to be chimerical. 23 Thus Quine, in one of his most pragmatic and behaviouristic moments, is ready to turn his back on Intentional idioms entirely, allowing them in his casual speech, but banishing them from the language of theory. 'One may accept the Brentano thesis as showing the

indispensability of Intentional idioms and the importance of an autonomous science of Intention, or as showing the baselessness of Intentional idioms and the emptiness of a science of Intention. My attitude, unlike Brentano's, is the second.'[24] This position, which is shared by behaviourist psychologists, is not merely the position that what you refuse to listen to cannot bother you. Our evidence that 'there really are' Intentional phenomena coincides with our evidence that in our ordinary language we speak as if there were, and if a science of behaviour could be successfully adumbrated without speaking as if there were these 'things', the insistence that there really are Intentional phenomena would take on a hollow ring. Behaviourism would attempt to discover extensional laws governing the occurrence of events (animal – including human – motions) that are *initially* given extensional, non-Intentional characterizations. If a truly predictive, extensional science of animal and human behaviour (specified in pure 'motion' terms and including all human verbal behaviour) could be produced, then the existence of Intentional idioms could be safely explained away as a peculiarity of natural languages, perhaps on a par with noun genders and onomatopoeia. Allowing working science to serve as ontological arbiter, one could claim that there really aren't any Intentional phenomena, and hence no science of Intention is needed.

Unfortunately for the alternative of behaviourism, however, so far the attempts to produce such an austere Stimulus-Response science have been notably unsuccessful. While behaviouristic research on animals and men

over the last several decades has been undeniably fruitful from the point of view of crucial data obtained, these gains have been achieved

independently of – and, in many instances, in spite of – the theories the experiments were intended to confirm or disconfirm. One could even

make a case for the claim that the value of experimental results has been in inverse proportion to the extent to which the shibboleths of orthodox behaviourism have been honoured. The difficulty the behaviourist has

encountered is basically this: while it is clear that an experimenter can predict rate of learning, for example, from the initial conditions of his mazes and the experience history of his animals, how does he specify just *what* is learned? It is certainly not the case that what the rat in the maze learns is a sequence of skeletal motions (for as Lashley's famous

experiments show, impediments in normal skeletal motion do not prevent

the rat from getting itself to the food). 25 Nor does it learn to move through a series of spatial juxtapositions no matter what. What it learns, of course, is *where the food is*, but how is this to be characterized non-Intentionally?

There is no room for 'know' or 'believe' or 'hunt for' in the officially circumscribed language of behaviourism; so the behaviourist cannot say

the rat knows or believes his food is at x, or that the rat is hunting for a route to x. The generalization of learning and the goal-directedness of the resultant behaviour have withstood all efforts to date to account for them as pure constructs out of the stimulus and response biography of the

animal, and the nature of the theoretical failures points to the possibility of a fundamental error in the approach. 26 The effect of these frustrations has been a relaxation of scruples, a tacit acceptance of Intentional

characterizations, so that interesting research can continue. 27 This strongly suggests, but does not prove, of course, that psychological phenomena

must be characterized Intentionally if they are to be explained and predicted, that no science of behaviour can get along without the

Intentional idioms.

What then of the alternative of Phenomenology, the establishing of ‘an

autonomous science of Intention’? 28 An Intentional science of behaviour would characterize the events of its domain in fully Intentional terms. Its programme would be to relate actions, beliefs, desires, intentions, rather than the supposedly ‘pure’ events of the behaviourists (stimuli and

responses characterized in extensional, ‘physical motion’ terms).

Explanations in such a science would characteristically take a form like

‘His desire to find shelter prompted him to try to find a way into the box’, and what is immediately apparent, but not as important as has sometimes been claimed, is that we *ordinarily* explain behaviour in the Intentional mode.

What is peculiar about such explanations is that they are not causal

explanations in the more or less Humean sense of the term. The key word in the example above is ‘prompted’; it is not to be replaced by ‘caused’.

The Humean doctrine is that causes must be identifiable independently

from their effects, for otherwise the statement of cause and effect will not be contingently, empirically true, as it must be, but analytic, i.e., true only in virtue of the meanings of the words. This independent identification and concomitant contingency is missing, however, when the antecedent is an

intention, the consequent an action, and the same can be seen to be true for all Intentionally characterized antecedents and consequents. 29 To see what this difference amounts to, consider first a faulty argument about a case of causal explanation, e.g., the claim that conception is the cause of

pregnancy. Hume requires causes and effects to be independently

identified, but *part of what we mean by* conception is that in the absence of interfering factors pregnancy results, so there is a *conceptual* (not merely contingent) connection between conception and pregnancy. Then

conception cannot be the cause of pregnancy. What is wrong with this

argument is that although conception can be characterized as what

pregnancy follows from, it can also be given independent

characterizations, in physiological terms, which make no mention of

pregnancy. If conception is *defined* as the cause of pregnancy, then it follows that 'Conception is the cause of pregnancy' is analytic, but it does not follow from this that conception is *not* the cause of pregnancy!

Provided there is a way of alternatively characterizing the event which is conception, it can be a perfectly good Humean cause.

The situation with Intentional explanations is different, however. It

follows directly from the Intentionalist's irreducibility hypothesis that no independent characterization of an Intentionally characterized antecedent is ever possible. To say that a particular Intentionally characterized

antecedent could be characterized in another way is to say that either the Intentional sentence announcing the occurrence of this antecedent has an extensional paraphrase (and this is ruled out *ex hypothesi*), or the Intentionally characterized antecedent can be given a different Intentional

characterization, but this is contrary to the fundamental principle of Intentionality, that Intentional phenomena are individuated by their

characterizations – a different characterization means a different

phenomenon. To take a particular example, consider a case where I

intentionally open a door and walk out of a room. Does my intention to

open the door *cause* me to open the door? If so, then we must be able to find another characterization for this intention, for as it stands it is certainly not conceptually independent of the consequent action. This

independent characterization would have to be either extensional – but we are supposing for the moment that this is impossible – or Intentional.

What other Intentional characterization of this intention could there be, though? The Intentionality of intentions is just that they have unique

characterizations; the intention to leave the room, for example, is not the intention to open the door, and so could not serve as an alternate

characterization of our initial antecedent. Nor can we avoid this dilemma by declaring that the cause of my opening the door was *not* my intention to open the door, but my intention to leave the room (which is an antecedent characterized independently of the consequent). For if I do leave the room, and do this intentionally, we shall have to find a cause for this as well, and if it is to be independently characterized it will have to be a yet more ulterior intention, say, the intention to see my brother, and so forth. It will not do for an Intentional science to try to get along in this way with an indefinitely large nesting of narrowly characterized actions caused by

more widely characterized intentions (opening the door being caused by

the intention to leave the room, leaving the room being caused by the

intention to see my brother, etc.) for if opening the door is an intentional action of mine, it follows that I must have had the intention to open the door *and* I must have opened the door *because* I had the intention to open the door. This must be the case since it is possible to intend to do X, and do X, and yet not do X intentionally, because one did not do X *because* of the intention. For example, an actress can intend to scream, actually

scream, and yet not intentionally scream, for though she did intend to

scream at the time, she actually screamed because she was genuinely

frightened. So the ‘because’ of Intentional explanations steadfastly resists treatment as a causal ‘because’; we must explain A’s intentional action X

by saying A did X *because* he intended to do X, and this intention cannot be given the independent characterization it needs to be a proper cause.

This impasse represents in itself an uncomfortable peculiarity for the

‘autonomous science’, but there is worse to come. The hegemony of

Hume-style causal explanation is not so secure that non-causal explanation counts as an overwhelming objection to Intentional science, but the closed nature of Intentional explanation has a further consequence that amounts to a radical asymmetry of our scientific world view – in many ways like the asymmetry feared by the identity theorists (see Chapter I). In the domain of causal explanation, that a particular event *a* is followed by another event *b* is explained by the invocation of some more general causal law, to the effect that all events of type A (which includes *a*) are followed by events of type B (which includes *b*), and this law may be explained in turn by being subsumed under or deduced from still wider

laws. In Intentional explanation, on the other hand, the sequences of events are so characterized that the occurrence of a particular consequent action is *explained by* the occurrence of a particular antecedent, say a perception or belief or intention, and there is no room for the question of why *this* consequent should follow *this* antecedent, and hence no room for any general law ‘explaining’ the sequence. For example, having said that my intention to leave was followed by my walking to the door, there is no

room for the question: why should *that* result (as opposed to, say, opening my mouth or raising my arm) follow the intention to leave. The ‘covering law’ to the effect that all intentions to leave are followed by walking to the door is silly and unnecessary; the occurrence of my walking to the door has *already* been explained by citing my antecedent intention.

In this way Intentional explanations assume the environmental

appropriateness of the connections between antecedent and consequent. 30

Thus there is a sense in which Intentional explanation is just the reverse of extensional, behaviouristic explanation. Behaviourism seeks to find

regularities and mechanisms that will explain the observed

appropriateness or adaptiveness of the connections between antecedent

and consequent events. Animal behaviour is generally appropriate to the environmental circumstances in which it occurs, and it is this ability to match behaviour to environment that the behaviourist tries to analyse by finding sequences of events that can be subsumed under general causal

laws.

For Intentional explanation, on the other hand, the fact that one event (as Intentionally characterized) is followed in an appropriate way by

another is not even contingent, and hence not subject to explanation. 31 The intention to raise one's arm would not *be* the intention to raise one's arm if it were not followed, barring interference, with raising one's arm, so the question of why one follows the other is superfluous.

It is this early end to explanation that puts Intentional science in

disharmony with the rest of science. As Pittendrigh observes, the

appropriateness or adaptedness of animal action implies organization

(which he distinguishes from mere order as being relative to an end). 'An organization is an improbable state in a contingent … universe; and as

such it cannot be merely accepted, it must be explained.'32 Thus the very feature which signals an end to explanation in the Intentional system

signals the need for explanation in the wider system of science as a whole.

The two sciences are not just separate, they are warring, for positions on what does and does not require explanation cannot be isolated within

autonomous branches of science. If adaptedness of animal behaviour

admits of, and requires, no explanation, then the improbable organization of which Pittendrigh speaks requires no explanation, and if this is so, we must either abandon the principle that the improbable requires explanation

– which would amount to the abandonment of the rest of science – or we

must maintain that such organizations are not improbable states of the

universe, which would require a total *bouleversement* of the physical sciences. The behaviourist has this much going for him: he is neither

anarchist nor revolutionary. The same cannot be said for the Intentionalist.

V THE WAY OUT

To sum up the results of the chapter so far, the effect of the Intentionality thesis is to give the old, ill-envisaged dogma that the mind cannot be

caged in a physical theory a particularly sharp set of teeth. The first challenge is the irreducibility hypothesis, that the Intentional cannot be reduced to the non-Intentional, or, as we have seen, the extensional. Then the evidence comes in that we can neither do without the Intentional, nor cleave to it alone, for there are signs that the possibility is remote of a successful non-Intentional behaviourist psychology; and the alternative of an entirely Intentional psychology would entail a catastrophic

rearrangement of science in general. This is not a formal dilemma, since on the one hand a forlorn hope may be held out that some future

behaviourist will be able to belie the many harbingers of doom and

produce a working non-Intentional theory, and on the other hand there are certainly some scientific revolutionaries who would relish a return to an anthropocentric and teleological world view at the expense of the current centrality of modern physics.

Fortunately, however, once the problem of Intentionality is clearly

expressed, it points to its own solution. There is a loophole. The weak place in the argument is the open-endedness of the arguments that no

extensional reduction of Intentional sentences is possible. The arguments all hinged on the lack of theoretically reliable *overt* behavioural clues for the ascription of Intentional expressions, but this leaves room for *covert*, internal events serving as the conditions of ascription. We do not

ordinarily have access to such data, so they could not serve as our ordinary criteria for the use of ordinary Intentional expressions, but this is just a corollary of the thesis that our ordinary language accounts of behaviour are Intentional, and says nothing about the possibility in principle of producing a scientific reduction of Intentional expressions to extensional expressions about internal states. Could there be a system of internal states or events, the extensional description of which could be upgraded into an Intentional description? The answer to this question is not at all obvious, but there are some promising hints that the answer is Yes.

The task of avoiding the dilemma of Intentionality is the task of

somehow getting from motion and matter to content and purpose – and

back. If it could be established that there were conceptually trustworthy formulations roughly of the form ‘physical state S has the significance (or means, or has the content) that *p*’ one would be well on the way to a solution of the problem. But if that is all it takes, the answer may seem obvious. Computers, we are told, ‘understand’ directions, send each other

‘messages’, ‘store the information that *p*’ and so forth, and do not these claims imply that some physical states of computers have content in the requisite sense? A hallmark of Intentional organisms pointed out by Taylor is that an Intentional description is one *for* the organism, for example the condition that is antecedent to intentional action is the condition of the environment as *seen* by the organism, but is it not true that the activities or motions of any cybernetic device are also relative only to the

environmental condition as ‘seen’ by the device? People who use

computers are accustomed to describing the operation of their devices in Intentional terms. If they are justified in speaking this way – and are not merely speaking ‘metaphorically’ – the Intentionalist claim will be

threatened, for then at least one sort of purely physical object will be understood as an Intentional system. It can be pointed out now that there is one serious flaw in our ‘hint’ however. A computer can only be said to be believing, remembering, pursuing goals, etc., relative to the particular interpretation put on its motions by people, who thus impose the

Intentionality of their own way of life on the computer. That is, no

electrical state or event in a computer has any intrinsic significance, but only the significance gifted it by the builders or programmers who link the state or event with input and output. Even the production of ink marks on the output paper has no significance except what is given it by the

programmers. Thus computers, if they are Intentional, are only Intentional in virtue of the Intentionality of their creators. People and animals,

however, are not designed and manufactured the way computers and their

programmes are, nor are they essentially in the service of interpreting, Intentional beings. (One could turn the argument around; then it becomes a rather top-heavy argument for the existence of an Intentional God – none of your theistic, abstract Gods – whom we are designed to serve.) If we are to avoid the God hypothesis, we must look elsewhere for a source of

Intentionality in living systems; we must find something else to endow

their internal states with content.

Following a well-beaten path, we can look to the theory of evolution by natural selection. The interpenetration of content and purpose has already been seen in the implication circle of belief and intention (see pp. 33–4), so it should not prove too surprising if the ability of the theory of natural selection to account for the apparent purpose-relativity of organs and

capacities of living things is also the ability to account for the content of certain of their states. Stronger links can be dimly seen. Intentional

description *presupposes* the environmental appropriateness of antecedent-consequent connections; natural selection guarantees, over the long run, the environmental appropriateness of what it produces.

An investigation of this avenue will take up the next few chapters. It can hardly be called striking out on a new trail. Considerable work has been done in what might be called the theory and construction of Intentional systems, but never to my knowledge has an attempt been made to spell out what the obligations and goals of this programme are. The burgeoning

fields of information theory and ‘artificial intelligence’ have produced a wealth of ‘models’ which may deserve to be called Intentional systems,

but the questions of whether or not these models do deserve this

appellation, and whether or not there can be *natural* Intentional systems along the lines of these models are questions to which little attention has been paid.

Theories of mind or behaviour in this general category are called

‘centralist’, in contrast to the ‘peripheralist’ theories of Stimulus-

Response behaviourism. While the peripheralist hopes to characterize

behavioural events and stimulation extensionally from the beginning, and arrive at extensional laws relating these, the centralist makes his initial characterization Intentional, describing the events to be related in law-like ways using either ordinary, or semi-ordinary, or even entirely artificial Intentional expressions. He then hopes that an adequate physical basis can be found among the internal states and events of the organism so that

‘reductions’ of Intentional sentences of the theory to extensional sentences of the theory are possible. 33 The ground rules for such ‘reductions’ have

not been set down, and this is one of the tasks of the next few chapters. A rudimentary excursion into neurology and information theory is

unavoidable, and both fields are jungles of conflicting claims and theories.

In part to avoid taking sides in these controversies, hypotheses will be put forward that ‘leave the details to the neurophysiologist’, and although

every care will be taken to provide that when this is done the hypothesis in question is compatible with whatever details the neurophysiologist might come up with, this is admittedly armchair science with its attendant risks.

The hope is that a strong case can be made for the *theoretical*

underpinnings of centralism, leaving as much room for empirical variation as possible. The examination of centralism will yield a number of

significant philosophical by-products, having to do especially with

consciousness, reasoning and intention, and these will be developed in Part

II.

3

EVOLUTION IN THE BRAIN

VI THE INTELLIGENT USE OF INFORMATION

In Chapter 1 we found a way of sidestepping the old and sterile problem of the ontological status of mental entities. In the place of an ontological division between phenomena or entities, we acknowledged only a division between the different things that we say, roughly characterized as a

division between the mental language and the language of science, or

physical language. In Chapter 2 this division was seen to coincide on a wide front, if not entirely, with the distinction between Intentional

sentences and extensional sentences, and this raised a fundamental

obstacle to our further efforts at relating mind to body, in the form of the Intentionalist thesis that it is logically impossible to ‘reduce’ the

Intentional mode of discourse to the extensional. Acquiescence in this

conclusion would leave large portions of our mental language discourse

inexplicable in terms of the physical sciences. Two attempts to get around the Intentionalist thesis were found unpromising. Attempts at a purely

extensional peripheralist science of behaviour have simply failed to

marshall their data into a working theory, and the failures bear all the earmarks of fundamental theoretical error; and an ‘autonomous science of Intention’ cannot co-exist with the rest of science. Since we apparently cannot do without the Intentional, and cannot allow it to remain

irreducible, the only course left is a more direct assault on the

Intentionalist thesis. The weak point in the arguments for the thesis was seen to be the reliance on overt, external behavioural cues as the

benchmark of extensional correlates. Would an examination of *internal* states and events gain us any leverage over peripheralist accounts and

allow us to prove the Intentionalist thesis wrong?

The theory that could do this would have to upgrade an extensional

account of the system of relations of internal, cerebral states and events into Intentional characterizations of these states and events, i.e., as events related to a content or message or meaning, events signifying or reporting or commanding. This is of course a standard practice of

neurophysiologists in expositions of their findings: they talk of neural signals, reports to the brain from the sense organs, and so forth, but this talk

is largely fanciful, and the rationale and justification of this step needs to be examined. What, if anything, permits us to endow neural

events with content? Can the rules governing this step of theory be generalized to allow us to speak confidently of neural events bearing

contents approximating to ‘the contents of our thoughts, perceptions and intentions’? To *begin* to answer these questions in this chapter we must venture into the area of neurophysiological hypothesis, stepping as lightly as possible, to see what the general shape of a theory would have to be to meet these requirements. That is, we shall investigate certain *minimum, necessary* conditions any centralist theory would have to meet, postponing until Chapter IV the question of whether a theory meeting these conditions has met the sufficient conditions for ascription of content to neural states or events.

We can call behaviour Intentional when it is of the sort that we normally characterize in Intentional terms, the sort that resists all efforts at extensional characterization. Thus searching for acorns and remembering to close the door are examples of Intentional behaviour, while stumbling, chewing and simply closing the door are not. We need not try to draw the line with precision since there are plenty of central cases we can consider before reaching any decisions about the penumbra, but as a general rule a bit of behaviour is non-Intentional if we could quite easily construct a device that performed it (a door-closer, a food-chewer), and is Intentional if it is not at all obvious that anything we might build could be said to be doing it (can we imagine a device which could be said, quite literally and unfancifully, to remember to close the door, to search for acorns, to

believe it is raining?). *Aficionados* of robots and those familiar with the claims made by workers in the area of computer simulation of behaviour

will perhaps reply that such devices already exist, but it is just these claims, among others, that we are scrutinizing. The controls and activities of computers can certainly be given an extensional description, and if they can also be characterized *justifiably* in Intentional terms we shall have *one* case of an Intentional-extensional reduction, and hence good reasons for expecting a similar reduction in the case of animals and people. The

strength of the analogy between human behaviour and computer behaviour

is thus a critical point which we will examine from a number of different points of view.

No creature could exhibit Intentional behaviour unless it had the

capacity to store information. For example, for a creature to exhibit

genuine goal-directed behaviour, the goals the creature had would have to

be ‘carried within it’ somehow, and ignoring animistic or mystical answers to the question how, the method of maintaining these goals within the

creature will have to be some form of storage in its material organization.

Moreover, the type of storage required must be what I shall call intelligent storage, the word ‘intelligent’ being used only as a tag for the time being, so as not to prejudge any questions about what constitutes genuine

intelligence. This notion of intelligent storage can best be made clear by the use of a few examples. Often when a computer is said to store

information the storage is nothing more than the capacity to produce a

sequence of characters in response to a particular cue. Thus one can store whole books in a computer memory and on giving the input, say,

‘Middlemarch’, one would receive as output the lengthy typing out of the novel word for word. A computer used this way is, of course, nothing

radically more than a tape-recorder with an automatic indexing system,

and its storage does not differ in type from old-fashioned library-shelf storage; only the mechanics of storage and retrieval are different. Neither the computer nor the library could be said in any sense to understand what was stored. Indeed this storage can be called information storage only by grace of the fact that the users of the output can interpret it as information.

One might speak of mountains storing geological and palaeontological information this way – all in precise sequence waiting to be interpreted.

Intelligent storage differs from this in that the information stored can be *used* by the system that stores it, from which it follows that the system must have some capacity for activity other than the mere regurgitation of what is stored. What counts as *using* the information is hard to say in many cases, but some computer programmes 'do enough' with the data they are fed to be strong candidates for the honour of intelligent storage.

To take a different example from the animal world, a parrot might have the ability to say 'fire hurts' and it might also exhibit fire-avoidance behaviour, but in the parrot's case we would not suppose there was any connection between the 'verbal' and non-verbal behaviour, unless, of course, the parrot, contrary to all we know about parrots, only spoke his little piece when the occasion called for just such a warning. The 'verbal' capacity of the parrot is a clear case of non-intelligent information storage, while his capacity to learn from experience in such a way that his behaviour improves in prudence is what I shall call the capacity for intelligent storage of information. The parrot, in learning to say 'fire hurts', does not store the *information* that fire hurts (at least it is not information *for* the parrot), even though we can imagine someone *using* the parrot – as one might use a writing tablet or tape-recorder – to store this information non-intelligently.

Non-intelligent information storage is nothing more than reliable plasticity of whatever lies between input and output, and hence we can see that the capacity for non-intelligent information storage must be the *basis* for the capacity for intelligent information storage. Typically, definitions of

information storage in the literature are definitions of non-intelligent storage, although they sometimes carry undesirable connotations of

intelligence. A definition of MacKay's, for example, is 'any modification of state due to information received and capable of influencing later

activity, for however short a time.'1 One must not read 'rationally influencing' or 'appropriately influencing' for 'influencing'; information is non-intelligently stored whenever the effect of an input is to contribute to the determination of a later output, *whatever this contribution is*.

We should reserve the term 'intelligent storage' for storage of

information that is *for* the system itself, and not merely *for* the system's users or creators. For information to be *for* a system, the system must have some *use* for the information, and hence the system must have needs. The criterion for intelligent storage is then the appropriateness of the resultant behaviour to the system's needs given the stimulus conditions of the initial input and the environment in which the behaviour occurs. Since

appropriateness is not an intrinsic physical or formal characteristic of any thing or event, no examination of the relations between intrinsic

characteristics of input and output will give us a clue about intelligence. A system that did not exhibit this capacity for environmentally advantageous response might be in fact a brilliantly conceived device for mathematical calculation (as could be determined by an examination of intrinsic

relations between its input and output), but it would be in the end only a tool; it would have no intelligence of its own, and would store no

information for itself. The capacity to store and use information

intelligently, then, does not emerge automatically at any degree of size or complexity of the information storage and processing mechanisms, but is an additional and separable capacity. The question now before us is what features a system must have if it is to acquire this additional capacity.

VII THE EVOLUTION OF APPROPRIATE STRUCTURES

The useful brain is the one that produces environmentally appropriate behaviour, and if this appropriateness is not utterly fortuitous, the production of the behaviour must be based somehow on the brain's ability to discriminate its input according to its environmental *significance*. If the brain cannot react differentially to stimuli in appropriate response to the environmental conditions they herald, it will not serve the organism at all.

How is the brain to do this? No physical motions or events have intrinsic significance. The electrical characteristics of an impulse sequence, or the molecular characteristics of a nerve fibre could not independently determine what the impulses *mean*, or what *message* the nerve fibre carries, and therefore what a stimulus – however complex – heralds cannot be a function of its internal characteristics alone. Therefore the capacity of the brain to discriminate by significance cannot be simply a capacity for the analysis of the internal structure, electro-chemical or cryptological, of the input sequences. It is easy to lose sight of this when we see how straightforward a task it is for researchers to determine the 'significance' of neural 'signals' in experimental animals. Whereas we, as whole human observers, can sometimes *see* what stimulus conditions cause a particular input or afferent neuron to fire, and hence can determine, if we are clever, its 'significance' to the brain, the brain is 'blind' to the external conditions producing its input and must have some other way of discriminating by significance. The criteria, for example, by which the MIT group[2] determines that certain afferent signals from the frog's retina signify sustained contrast or moving edges or convexity cannot be used by the brain of the frog to discriminate these signals, because the frog's brain cannot observe the frog's retina, cannot tell where these signals are coming from.

Since environmental *significance*, even in the attenuated sense in which retinal impulse streams *signify* certain retinal conditions, is not an intrinsic physical characteristic, the brain, as a physical organ, cannot sort by significance by employing any physical tests. The only other

explanation that would be acceptable to the physical sciences is that the brain's capacity to discriminate appropriately is based on chance. That is, a particular pathway through the brain might just happen – entirely

fortuitously – to link an afferent (input) event or stimulus to an efferent (output) event leading to appropriate behaviour, and if such fortuitous linkages could in some way be generated, recognized and preserved by the brain, the organism could acquire a capacity for generally appropriate

behaviour.

Let us mean by a *functional structure* any bit of matter (e.g., wiring, plumbing, ropes and pulleys) that can be counted on – because of the laws of nature – to operate in a certain way when operated upon in a certain way. Obviously just about anything can be viewed as a functional structure from one point of view or another. A functional structure can break down –

not by breaking laws of nature but by obeying them – or operate normally.

A nail is a functional structure; so is a gall bladder, and an open telephone line between Washington and Moscow. Given a brain with an initial

plasticity or capacity for producing different functional structures as a result of input, the key to utility in the brain must be the further capacity to sort out these functional structures, keeping and using those that are useful to the survival and comfort of the organism, and eliminating or

refraining from using the harmful ones. We cannot suppose that harmful

structures suffer in themselves from any physical defect – e.g., chemical instability or a tendency to atrophy – nor that useful structures are

particularly robust, so if we are to have an analogue of natural selection do the sorting of structures, it cannot operate on a principle of physical fitness; nor can it be, as we have seen, that useful structures are useful in virtue of any distinguishing intrinsic physical characteristic that could be keyed on by a sorter. There are other ways of establishing a sorting

principle, however, and an externally grounded sorting mechanism that

meets the requirements we have enunciated can be described with the help of a very elementary excursion into a hypothetical evolutionary history.

At a very early point in evolutionary history, organisms appeared with

simple nervous systems; contact with their surfaces produced electrical activity similar to that of neurons. The value of this phenomenon

depended on the result it happened to trigger. Suppose there were three different strains of a certain primitive organism in which a certain

stimulation or contact caused different ‘behaviour’. In strain A the

stimulation happened to cause the organism to contract or back off; in

strain B the only behaviour caused by the electrical activity in it was a slight shiver or wriggling; in strain C the stimulation caused the organism

to move towards or tend to surround or engulf the point of contact causing the stimulation. Now if the stimulation in question happened to be caused more often than not by something injurious or fatal to the organism, strain A would survive, strain B would tend to die off and strain C would be

quickly exterminated (other conditions being equal). But if the stimulus happened to be caused more often than not by something beneficial to the organism, such as food, the fates of A and C would be reversed. Then,

although all three responses to the stimulation are blind, the response that *happens* to be appropriate is endorsed through the survival of the species that has this response built in. This observation taken one way is

tautological; what is appropriate tends (by definition) to aid survival; what is inappropriate tends (by definition) to kill off the organism. The species that survive are the species that happen to have output or efferent impulses connected to the afferent or input impulses in ways that help them survive.

As the evolutionary process continues, the organisms that survive will

be those that happen to react differently to different stimuli – to

discriminate. Thus if strain A backs off for both stimuli X and Y, while strain B backs off for X and approaches for Y, and if X happens more often than not to announce injury and Y happens to be caused more often than

not by nourishment in the environment, strain A will die of starvation

since it runs from both danger and food, while strain B will survive by discriminating appropriately. The discriminatory behaviour of strain B is only blind, dumb-luck behaviour; that is, it is the fortuitous and

unreasoned result of mutation, the appropriateness of which is revealed by the survival of the strain. In this way a variety of simple afferent-efferent connections can be genetically established, and once they are firmly

'wired in' the afferent stimuli can be said to acquire a *de facto*

significance of sorts in virtue of the effects they happen to have, as

stimuli-to-withdraw-from and stimuli-to-remain-in-contact-with.

Moreover, natural selection ensures that the former will be in fact danger signals and the latter, beneficence or security signals – harbingers of good in one respect or another. Of course nothing in the organism will *recognize* these stimuli as danger or security signals, unless one wants to say that the organism's good fortune to be so wired as to react appropriately to these stimuli amounts to its recognition of their import, but this would surely be an overly fanciful way of speaking.

So far so good: natural selection can provide for the dullest sort of appropriate reflex responses to stimuli discriminated by their meagre, in

fact binary, ‘significance’. Other genetically grounded connections besides those rudimentary arcs would be possible and in fact likely. In all species the pain network is at least to some extent wired in (and we shall see later why this must be so), and the transmission of the controls of rigid

‘instinctual’ behaviour must also be genetic. Any afferent-efferent

connection that was regularly appropriate would have survival value, the likelihood of survival depending on how regular the beneficial

environmental results of the response motion are. It is presumably

possible in principle for evolution to produce an organism with a useful brain that was entirely genetically pre-wired in this way and had no

plasticity at all. This would depend first on the genes’ having sufficient information-transmission capacity to transmit complete wiring-diagrams

from generation to generation. Such an organism would ‘know it all’ from birth and be unable to learn – not that it would need to. This could only happen where the environment in which the species lived consisted of

utterly stereotypic situations and remained extremely uniform throughout the aeons of evolution. Only where the appropriate response to a stimulus remains unchanged from individual to individual and generation to

generation would pre-wiring on such a scale have any survival value. And of course no matter how precocious the organisms would appear in their

natural habitat, in an alien environment they would be worse than

moronic. Such rigid behaviour patterns, or tropisms, are of course

common among insects and other lower animals, and, for example, if fires became regular features of the environment of the phototropic moths, they would soon become extinct. 3 Thus a preponderance of tropistic behaviour controlled by pre-wired afferent-efferent connections can become an

evolutionary trap for a species if the environment changes.

If too much inherited wiring is a bad thing, a certain amount is

absolutely essential. Nothing about an afferent impulse *by itself* could mark it as positive or negative ‘feedback’ and thus start the learning

process. Afferent impulses alone could have no useful bearing on the

behaviour of an organism, so there is no hope of achieving utility unless some afferent impulses are pre-wired to the appropriate responses. The

problem then becomes: how does the pre-established ‘significance’ of

some afferent impulses allow the brain of a learning organism to

discriminate appropriately the *other* impulses, which are not genetically endowed with any ‘significance’?

A fairly common picture of the brain that might suggest itself in

response to this question must be rejected now. This is the picture of the brain as composed at birth of two sides, afferent and efferent, with a few pre-wired connections between the two sides (reflexes and tropisms), but the rest of the gap free of connections, awaiting the dual gift of efferent coordination and afferent analysis, before the afferents are connected to the ‘right’ efferents. This view may be a hangover from the telephone

switchboard motif of a few decades ago, which made it comfortable to

envisage two essentially separable switchboard systems: the afferent caller announces his business and the operator plugs him in to the appropriate efferent receiver. This is a hopeless way of looking at the brain, since it still requires ‘the little man in the brain’ who understands, reasons, and in general intelligently uses the brain, thereby robbing the brain of just those intelligent functions we are trying to endow it with.

The insoluble problem of getting the stimulus to *find* the right response, of making the right connection across the great divide, can be avoided

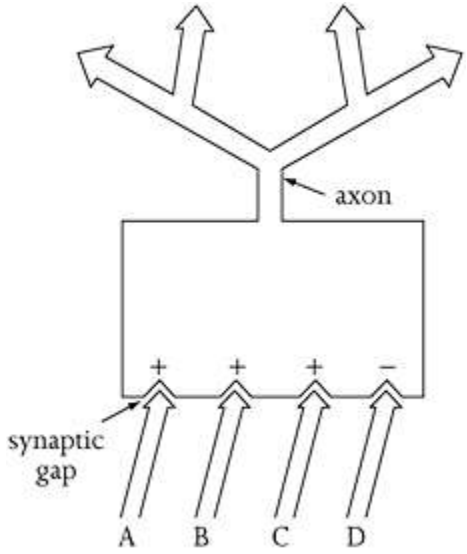
only if it is supposed that the afferent and efferent sides of the brain are richly, if to some extent randomly, interconnected from birth. The classic function of natural selection is to cull repeatedly the few good from an abundance of candidates, and if the process of evolution is to be brought into the brain, there must be an initial abundance from which to cull the

survivors. Skinner has the concept of ‘operant’ behaviour, which is not stimulated or ‘elicited’ but just ‘emitted’ by the brain – apparently by the efferent side of the brain acting alone. 4 The meaningless babbling of an infant and its apparently random limb movements are examples of operant behaviour, and Skinner holds that somehow operant behaviour can be

refined and connected to a stimulus cue. Thus the child learns to speak and walk. Skinner’s problem is how to make the afferent stimulus cues jump

the gap to the efferent side and select the appropriate ‘emissions’. If instead of supposing with Skinner that the apparently random operant

behaviour is in fact randomly emitted by the efferent side, we suppose that it is stimulated – entirely inappropriately – by the as yet unstructured and unanalysed afferent barrage, the problem is no longer how the afferents get to their appropriate efferents, but how the appropriate interconnections among the many inappropriate ones get weeded out for survival. What is



needed is some intra-cerebral function to take over the evolutionary role played by the exigencies of nature in species evolution; i.e., some force to extinguish the inappropriate. A capacity for propagation is also needed to provide continued abundance for intra-cerebral selection. In inherited pre-wiring we have the basis for such capacities, but in order to explain how this might work we must delve deeper into the physiology of the nervous system.

The nervous system is composed of two major types of cells, neurons

and glial cells. The glial cells are generally supposed to have the function of providing life support for the neurons, but perhaps they also participate in the functional plasticity involved in information storage. Whether they do in fact have this latter function is immaterial here. The neurons, which number in the neighbourhood of 10,000,000,000 in the human brain, are

the transmission and switching elements of the brain, and may contain

within themselves the whole capacity for information storage, leaving the glial cells to their more mundane role.

Schematic Diagram of a Neuron

If the neuron has a threshold of + 2, it will fire impulses along its axonal branches only when it receives impulses simultaneously from at least

two of A, B, C, but not D – or all four.

Each neuron has an input end, consisting of many terminals to which are led the outputs from other neurons or, in the case of neurons on the

periphery of the nervous system, from receptor cells in sense organs. The

neuron has a single output line, the axon, which branches after leaving the cell body into many outputs which lead to the input terminals of other

neurons. The endbulbs of the axon branches do not quite touch the input knobs on the receiving neurons; the gap between them, or synapse, is

crossed only when the impulses arriving at the synapse achieve a certain minimum frequency. The millions of neurons, particularly the afferent

neurons, are arranged in regular ranks, so that all the output branches of neurons in one rank connect to inputs of neurons in the next rank up. There are important exceptions to this directionality, such as the ‘descending effects’ that seem to be critically involved in the process of perceptual analysis, but they do not concern us here.

Of paramount importance to the theory to be proposed is the

phenomenon of *threshold*. Some synaptic crossings contribute to the excitation of the neuron and some inhibit its excitation. Each neuron has a

'statistical' or threshold mechanism so that it fires its output only when the weight of excitatory crossings at a given time exceeds the weight of inhibitory crossings by a certain value. To simplify for our purposes, if each excitatory crossing is given a weight of +1 and each inhibitory

crossing a weight of -1, a neuron with an excitation threshold of 2 would fire its output only when, for a short moment, the sum of all crossings >2.

The threshold of a neuron is *variable*. Frequent firing of a neuron tends to lower its threshold while inactivity raises the threshold.

This much seems quite well established by the neurophysiologists

although the importance and roles of these features are widely debated. It is in any case enough for the hypotheses we need, and in some ways more than enough. All we need is a multitude of switching elements arranged

with enough directionality to allow us to speak in a general way of higher and lower levels, and a general rule (though it need hold only over a

certain range of values, and need not be unexceptioned) that the firing of a switching element increases the likelihood of its firing again. This last condition, for which there might be equally suitable analogues, gives us our principle of 'species' propagation.

We need not reach any decisions regarding the disagreements in the

field over the mechanics of cerebral plasticity or transmission. The role of RNA changes in the cells, the chemistry of the synaptic crossing, the

generalized effects of drugs on plasticity, need not concern us provided that our general features and principles are embodied one way or another

in the brain. There are other elaborations and complications of this picture that are more relevant to our general outline, but in the interests of

maintaining our conceptual account as much as possible invulnerable to

empirical disconfirmation in the laboratory, these will be ignored. For example, it is tempting to suppose, and there is some evidence for

supposing, that particular synapses that regularly contribute to firings of a neuron tend to lower their frequency requirements, perhaps accomplished by a narrowing of the synaptic gap due to stimulated growth of the

endbulb and dendritic knob, but it is not essential that we suppose this.

There is one other feature of the brain's physiology which it is important to mention, not because it is a feature required by our particular

hypotheses, but because it is a feature required of all reliable information processing systems, and that is the brain's use of redundancy and the

'ambiguity' of neuronal 'signals'.

No sense has yet been given to the claim that a neuron's impulses are

signals with content or meaning, but if, for example, a particular neuron in the optic nerve fires its output if and only if there is a particular pattern of stimulation on the retina (due to the particular summing effects of the neurons in the lower ranks leading to its input), in a borrowed sense one could say that the neuron's output is unambiguous. However, except at the most peripheral levels, neuronal firings turn out to be, in this sense, ambiguous. That is, a wide variety of different, in fact very dissimilar, stimulus patterns may cause a neuron to fire, so that its signal is highly ambiguous. This fact, distressing to the neurophysiologist intent on

'breaking the neural code', is vital, however, to the successful functioning of the brain. The brain, for all its occasional lapses, is a highly reliable organ; seldom if ever does a complete failure of stimulus interpretation occur. If each neuron had only a function, and this function was not

duplicated by other neurons, the death or malfunction of any neuron would throw all that followed into disorder. At the peripheral level – near the retina, for example – the death of a neuron might only cause a small ‘blind spot’ or imperceptible loss in colour discrimination or something of the sort, but if a single neuron at a high level were to carry single-handedly some information about a highly complex pattern of stimulation, its

breakdown would cause something like total blindness for particular

shapes or wildly mistaken identification of objects in the visual field.

Neurons do not regenerate like other cells, and their mortality rate may be

in the neighbourhood of one neuron a minute. Neurologists estimate that random malfunction of about one per cent of the neurons in the operation of any brain tissue or structure is normal. Clearly the reliability of the brain is greater than that of its components. Arbib presents a calculation to show the effect of random failures:

Consider a chain of *n* modules [neurons] and assume that there is a probability *p* of malfunction for each neuron. Then the probability that the output of the chain is correct is, to a first estimate, (I - *p*) *n*.

Now no matter how small *p* is, (I - *p*) *n* gets to a ½ when *n* is made large enough and if our output is equally likely to be right or

wrong, it is of no use to us! 5

He goes on to point out that if *p* is one per cent, a neuronal chain of only 70 elements will have a probability of correctness of ½, and 70 elements is not very deep for the human brain, with its 1010 neurons.

Reliability of transmission using unreliable elements can be achieved

provided there is signal duplication in some form. If, for example, a

‘message’ is transmitted simultaneously by five neurons, and the

probability of successful transmission for each neuron is high, say 0.99, the probability that successful transmission will occur in at least three channels is much higher. Then, if a statistical or vote-taking mechanism is inserted between each level of transmission, random errors due to

malfunction will be absorbed as soon as they occur. The variable threshold capacity in the neuron can perform this function, provided the redundancy of signals is great enough, and provided there is a rich enough

interconnection of outputs with next-level neurons.

Simple redundancy, however, with each neuron's output serving one

purpose, would require an inefficient multiplication of elements. If, on the other hand, the signals fired by each neuron are ambiguous (as they are), if each neuron contributes to many different multiple transmissions,

redundancy can be achieved with less elements. It is then the more or less simultaneous concatenations of neuronal outputs or signals that are

unambiguous, rather than the outputs of individual neurons. The

convergence of different concatenations of ambiguous signals at each

succeeding level would partly resolve the ambiguity just as the

convergence of ambiguous definitions determines unique or nearly unique solutions to crossword puzzles.

The crucial point that emerges from this is that the candidates for

vehicles of content or significance in the brain are *compound*. Afferent-efferent functional structures, which are to be sorted according to their appropriateness, have parts and could be 'rebuilt' piecemeal under certain conditions. The features of variable threshold and compound 'signals',

together with the hypothesized initial situation in the brain of rich

afferent-efferent interconnection and some partial appropriate pre-wiring, provide the elements needed for a hypothesis of evolution in the brain

capable of explaining the brain's ability to discriminate by significance and store and use information intelligently.

The problem set up earlier was how the brain could cull out the

appropriate afferent-efferent connections from the initial abundance of haphazard connections, but given the compound nature of neural signals

we can no longer look for there to be whole compounds among the

initially senseless fabric that are appropriate. That is, the odds are

certainly against finding fortuitous structures sufficiently large and

complex to produce or direct anything as sustained as a bodily motion

would have to be, to be a demonstrably appropriate response to the

stimulus environment. Where no connections would qualify as

appropriate, how is selection to proceed?

The answer comes from taking another hard look at evolution of

species. For there to be evolution, there must be conflict between some features in the environment and the species to be eliminated. The only way any functional structures could be sorted within the brain would be if some of them were to conflict with the pre-established, wired-in, appropriate connections. There must be conflict and something must give. Clearly

what must stand firm are the inherited connections. No other conflict, and no other outcome of the conflict, would resolve itself along appropriate lines. The inherited wiring or programming must be granted hegemony in

all conflicts if the plasticity of the brain is not to undo the work of species evolution and leave the animal with no appropriate responses at all.

So long as the other initially salient neural pathways are haphazard and uncompounded into ‘coordinated’ functional structures, what sense could be given to a notion of appropriateness or inappropriateness of these

connections? All structures, it would seem, are going to be equally neutral

in this regard. One bit of babbling or finger-twitching is no more or less appropriate than another. Some such structures, however, might conflict internally with the pre-wired connections, and although these would not be environmentally inappropriate structures by themselves, they would stand in the way of the completion of the pre-wired connections. These inherited links must, in addition to stimulating certain muscles in a certain sequence when presented with certain stimuli, also block the stimulation of

conflicting muscle motion and the perpetuation of any neural structures that would in any other way interfere with the operation of the inherited links. Since for any afferent-efferent functional structure to become

genetically established it must be environmentally appropriate over the long run, and since for any such structure to be appropriate it must be capable of surviving in a plastic brain, all genetically established afferent-efferent structures must have, in addition to the appropriateness of their unimpeded function, the general capacity to inhibit competing

connections. If this is the case, any of the initially haphazard connections that inadvertently competed with pre-wired connections would be

inhibited and eventually, through inactivity, become inoperative, while any compatible haphazard connections would be allowed to complete

themselves, and, by our principle of propagation, they would tend to recur.

Just as in species evolution, it is thus not death itself that extinguishes a species, for all animals die and all neural events come to one end or

another, but the *failure to reproduce*. This would have the effect of pruning the initially unstructured connections along lines at least

compatible with and occasionally contributory to the appropriate inherited links already endorsed by species evolution. It would allow for the

reproduction of everything that is at least not inappropriate, taking

whatever inherited links there are as the arbitrary but contingently

accurate standard of appropriateness.

Harlow uses the term ‘baroque’ to describe those features that become

genetically established through natural selection and exceed the

functional, and this capacity for the propagation of the baroque is essential to the evolution of many capacities found in nature. Wings, for example, could not evolve fully developed in one fell swoop, and yet until they are fully developed, they have no positive survival value. The ability of a species to maintain through generations a fractionally or potentially

appropriate feature is the *sine qua non* of complex capabilities and

structures, and this holds particularly true for the obviously enormously sophisticated structures that must be required to control the behaviour we observe in animals and human beings.

The gradual effect of this gentle sorting action will be to establish new functional structures, but if these are to have any permanence they must similarly be capable of overruling competitors, although here which

competitors overrule which will not be entirely a matter of precedence of establishment, as it is with the pre-wired structures, for we want to allow for the unlearning of behaviour that eventually turns out to be

inappropriate. As relatively permanent new structures are laid down, the efficiency of the sorting action will, of course, increase. An early

manifestation of this evolutionary pruning will be the gradual smoothing out of bodily motions into more coordinated and graceful motions, and the

resulting locomotion, the capacity for which is built up piecemeal, will bring to the animal new ‘experience’ in the form of novel stimulus

patterns. These in turn will ensure that a constantly changing and novel afferent input will be presented to the brain (the analogue of mutation in species evolution) and the efferent continuations that happen to result from these new afferents will in turn be sifted. The effects of increasingly appropriate motion include an improvement in the quality of information brought in by the afferent barrage, as appropriate efferent structures

controlling the focusing of attention, opening the eyes, and so forth

become established. (There is a good deal of evidence that the controls for maintaining steady eye position, focusing, and maintaining standard

orientation of retinal images are genetically transmitted, thus ensuring from the very beginning some regularity in the afferent barrage from the eyes, but the extent of this is not important here.) Thus the process is a repeated self-purification of function, gaining in effectiveness as more and more not inappropriate structure becomes established.

Intuitively, the speed at which the evolution takes place will depend in part on the extent and rigidity of the initial programming or pre-wiring, and this is borne out in nature. Many animals are born with mature

capacities for locomotion and discrimination of objects in their

environment, but the greater the initial ability, the more rigid the brain, and hence the less adaptable the animal. More intelligent animals require longer periods of infancy, but gain in ability to cope with novel stimuli because of the higher proportion of ‘soft’ programming – programming

not initially wired in and hence more easily over-ruled by novel stimuli.

The speed of evolution is in any case incomparable to the speed of species evolution, for the counterparts of generations endure not for decades or months or even minutes, as in the case of some primitive organisms, but for a few milliseconds.

If it be doubted that such a slight force could account for the learning capacities of animals and men, the fact that species evolution has

produced ‘instinctual’ behaviour in some animals the equal of learned

behaviour in others is some support for the claim. Babies must learn to see and walk, but whatever controls this in babies has a counterpart in chicks and puppies, and in these creatures the controls are clearly almost entirely inherited. Species extinction is as slight a force as the extinction through incompatibility posited for the learning brain, and yet species extinction has been a strong enough force over the years to produce such complex

behavioural controls as those governing the ‘territorial’ behaviour of some birds, food discrimination, and specific patterns of defensive behaviour.

Implicit in these arguments is a corollary to the effect that the

Lamarckian hope that some acquired characteristics may be genetically

transmitted is gratuitous. The fear that makes Lamarckian hypotheses

attractive is the fear that species evolution by itself would not be effective enough to produce the sophisticated ‘instinctual’ behaviour observed in the animal kingdom, so individual acquisition of know-how is rung in to help. But I have argued that only an intra-cerebral evolutionary process could account for such individual acquisition, and if an intra-cerebral evolutionary process *can* produce sophisticated behaviour, it follows that over a longer run species evolution can do the same. The transmission of acquired characteristics is not ruled out by this argument; it is just denied the crucial role it might seem to have.

The intra-cerebral evolution hypothesis also allows the controversies

over instinctual behaviour and the interpretation of deprivation

experiments to be seen in a new light. 6 The standing difficulty with deprivation experiments has been the near-impossibility of so reducing the

stimulus environment of the animal from birth that the possibility that the behaviour in question is learned can be ruled out. This has been a

difficulty because some stimulation is always necessary just to 'trigger'

the behaviour in question. The results of experiments have tended to blur the fine line between innate and learned behaviour that was seen as a

desideratum. Does the animal have the particular behavioural capacity intact at birth, or does it have some inner state at birth which allows it to

'learn' the behaviour almost instantaneously when the right stimuli are present? The distinction loses much of its importance given the inchoate view of learning presented here. The existence of some degree of wired-in behavioural controls is established here not by the results of deprivation experiments but on conceptual grounds alone. That is, it is argued that without some such foundation for appropriate behavioural discrimination in the brain, the brain as a physical organ could not learn at all, since it would have, as it were, no 'standpoint' from which to make initial

discriminations. The extent of pre-wiring in each species is subject to experimental determination, but there is no assurance that the exact limits of pre-wiring will be determinable via behavioural manifestations. That is, such basic pre-wired controls as those governing reflex withdrawal from painful stimuli have obvious behavioural manifestations, but much more

sophisticated behavioural controls may be genetically transmitted and yet because they are only *partial* in their pre-wired form, deprivation experiments would not reveal their existence. It is likely, on this view, that partial appropriate afferent-efferent connections could be established the completions of which would have to be learned. There is no sharp

distinction in efficacy between species evolution and individual intra-

cerebral evolution, so where species evolution leaves off and intra-

cerebral evolution takes over is a matter of no great importance as far as the survival value of the pre-wiring goes. A set of *potential* behavioural

controls has virtually the same survival value as complete behavioural

controls, given a regularity in the early learning environment of the

species ensuring the completion of the controls in most cases. Behavioural evidence for such partial structuring of the infant brain might be

extremely indirect and not at all conclusive. For example, it might well be the case that in a human being there is a partial smile-discrimination

mechanism, the completion of which requires a relatively large amount of learning, including at the very least the development of locomotion and sensory discrimination in general. The evidence for this is quite tenuous.

Babies seem to respond appropriately to smiles very early, which is

remarkable considering the complexity of what is communicated by a

smile and the paucity of corroborating evidence in the environment for a smile's significance. There can be nothing intrinsically friendly about the

spatial configuration of a smiling face, and furthermore there must be a remarkable lack of uniformity in the retinal projections of different

smiles. The universality of significance of smiles among the people of the world contributes to the suggestion that there is some partial

discrimination mechanism genetically transmitted, and of course such a

mechanism would have survival value since the early recognition of, say, parental approval or disapproval is a valuable capacity in the learning child – if not today, very likely in primeval days when an unlearned lesson could be fatal. (Survival value would depend, of course, on a concomitant inherited tendency to smile when one wished to show approval, pleasure, friendliness, etc. Alternatively, an inherited smile-recognition system could be entirely baroque or a no longer useful relic from man's simian past.) Such a partial pre-wiring, if it exists, would not even come into play, would have no behavioural manifestations at all, until considerable

learning had occurred, and hence would be a bit of inherited behaviour

control quite inaccessible to testing by deprivation experiments.

This advance in outlook for deprivation experiments is just a special

case of a general advance in outlook for behavioural theory provided by the evolutionary hypotheses sketched in this chapter. The difficulty with behaviourism is, tautologically, that one's subject matter is limited to behaviour, and the difficulty with this is that behaviour does not allow itself to be divided into the right sorts of parts. The strong point of S-R

behaviourism is its recognition, in vague intuitive form, that there must be something in the nature of a carrot and stick, reward and punishment,

survival and extinction, if learning is to be explained. Somehow sensory

'feedback' must be distinguished as at least positive or negative if the creature is to make any headway at all. And, since there is nothing

intrinsically positive or negative in any stimulus, there must be something like the evolutionary conflict sketched here. But so long as the

evolutionary conflict is dogmatically asserted to be entirely *overt*, manifested in trial-and-error behaviour – and this is what is involved in the behaviourist creed that behaviour must be a function of past behaviour and stimulation – the evidence will just not support the theories proposed.

Animals do not engage in enough trial-and-error behaviour to ensure the development of their behaviour along appropriate lines. Except at the

lowest evolutionary levels of life, where S-R behaviourism fits the facts quite well, animals need not in every case run their behaviour all the way

to painful consequences before learning that it is inappropriate. So long as it is behaviour, and not behavioural controls in the form of afferent-efferent interconnections, that is deemed to be reinforced or extinguished, one's theory is stymied by the fact that one cannot divide behaviour into fractionally appropriate, fractionally 'rewarded' bits and fractionally

inappropriate, fractionally ‘punished’ bits. The absence of overt reward in the form of, say, food, and the absence of overt punishment in the form of pain, in cases where the experimental animal *approximates* appropriate behaviour or partially completes an appropriate or inappropriate response has led to the postulation of such theoretical monsters as ‘fractional

anticipatory goal responses’, ‘expectancies’ and ‘partially reinforcing stimuli’. 7

Once one recognizes the need for a carrot and stick in learning, pleasure and pain stimuli offer themselves as the obvious candidates for these

roles, and at a low enough level they prove adequate to the task. But when the appropriate behaviour requires extended motion and control, pain and pleasure do not suffice as directors since their force cannot be transmitted from the *whole* behavioural response to its parts. How can ‘a step in the right direction’ or ‘a step in the wrong direction’ be recognized by the organism if it is not immediately rewarded or punished? A covert, internal carrot and stick must do the job, and here the impossible notions of

fractionally appropriate and fractionally inappropriate motions can be

replaced by the notions of afferent-efferent functional structures either compatible or incompatible with the overruling pre-wired structures. A

‘fractionally inappropriate’ structure could be discriminated and

extinguished not by virtue of any overt semi-punishing stimulus but by

virtue of its being blocked by internal programming. Trial and error *must be* required for learning, but there is no reason why it must all be external and overt, all in terms of behavioural trial and error.

VIII GOAL-DIRECTED BEHAVIOUR

The principles of evolution proposed to explain learning and

discrimination in the brain have the capacity to produce structures that have not only a *cause* but also a *reason for being*. That is, we can say of a

particular structure that the animal has it *because* it helps in certain specified ways to maintain the animal's existence. It is a structure *for* discriminating edible from inedible material, or *for* finding one's way out of danger.

Here we must be careful to distinguish between there being a reason for the animal's having a structure and the animal's having a reason for having the structure. Even a human being could not normally be said to have a

reason for having a particular neural structure (although a human being might, for example, have a reason for having an electrode implanted in his brain). Thus when there *are* reasons for the presence of neural structures, it is not the case that the animal or person *has* these reasons. Nor is the *raison d'être* of a neural structure just like the *raison d'être* of, say, a can-opener, for a can-opener's existence depends on the recognition of its

raison d'être by its maker, whereas no one and no thing (e.g., Nature) need recognize in any way the *raison d'être* of a neural structure for it to exist for the reason it does. The cash value of saying that a neural structure exists for a reason is just this: were the necessary conditions for the survival of a particular animal and the environmental circumstances in

general other than they are, such that the neural structure in question would not have the role in survival it has, the structure would not exist. 8 It is through observing and evaluating circumstances and behavioural

efficacy that we are able to frame the *raison d'être* of a structure, but no such activity of observation and evaluation is a pre-requisite for the

appearance of the structure on the evolutionary scene.

Although we have found structures with a reason for being, the only

control structures described have been of a stimulus-response variety,

granting that no limits have been set on the mediation between stimuli and responses, and a centralist account of learning has replaced the

peripheralists' external rewards and punishments. Critics of stimulus-response theories have often contrasted response behaviour with goal-directed behaviour. Can these structures that have themselves a reason for being provide for goals, for an animal's having a goal and hence having a reason for doing something? To answer this question we must first establish criteria for goal-directed behaviour.

'Goal' is a popular word among computer programmers. Every programme has a goal, and some have hierarchies of goals. A great deal of this 'goal' talk is exaggeration. When a programme is designed merely to terminate in the solution of a single problem for which a linear enumeration of steps is provided by the programmer, it is stretching a point to say that the computer – or the programme – *has* a goal; the programmer himself has the goal, and the programme is merely his means of achieving it. Other programmes, however, are designed to accept a variety of quite different problems, and once a 'goal' has been set for it (e.g., solving Problem P) it proceeds to 'search' for a solution with no further direction from the programmer. Here it looks much more plausible to say that the computer has been given a task to complete, so that it is the computer itself that (at least momentarily) *has* and *is directed by* a goal.

For example, Newell and Simon, in describing their GPS (General Problem Solver) computer programme, 9 claim that GPS operates in a goal-directed way. GPS consists of a number of 'sub-routines' which perform mathematical transformations on input expressions, together with an ordering system that allows the computer to move from sub-routine to sub-routine. Giving GPS a problem to solve involves specifying an input expression and an end-state; typically the input expression is a set of

premises and the end-state is the last line of a proof to be constructed, and thus the goal of the computer's activity is said to be embodied in the

specification of end-state, and the means to the end are the series of subroutines, which the computer tries one after another. The result of a

completed sub-routine is checked by the computer against the specified

end-state. Partial similarity is treated as apparent progress towards the goal and the result is saved for further transformation. Does this system of subroutines and specified end-state provide an adequate model of goal-direction?

It is particularly interesting to note that the authors of GPS are prepared to call such activity goal-directed in spite of their acknowledgement that not much in the way of heuristics is built into their programme for

changing the order in which subroutines are tried or for ruling out

obviously inapplicable subroutines. The computer rather inelegantly

grinds away until the end-state is achieved or it runs out of sub-routines.

GPS is not very insightful about recognizing progress; any similarity,

however unpromising to human observers, is treated as apparent progress towards the end-state, deserving further work. At first this may strike us as a serious shortcoming. One is reminded of the extremely vague but

compelling notion we have of a goal, lighting the way, informing our

choices, hovering and helping us decide as we pick our way towards it.

There seems to be little or no *direction* in GPS, and we might decide to call its activity an example of goal- *terminated* behaviour, but not goal-directed behaviour. If we are to make this distinction, however, we must be prepared to provide a more exact description of genuine goal-direction.

We speak of goal-directed behaviour in animals, but when we do our

standards are set quite low. *Prima-facie* evidence for goal-directedness in animals is the production of a repertoire of alternative motor patterns until one ‘attempt’ pays off by bringing about a certain sensed environmental end-state. The more ‘random’ the successive attempts appear – especially if dead-ends are repeated – the less we are inclined to call the behaviour goal-directed, but there are no minimum standards of elegance in motor

choice. We cannot even rule out as goal-directed the repetition of deadends, for even fully rational human beings often make repeated attempts at unlikely means to their ends. Are we to say that the prisoner who tries repeatedly to scale the unscalable prison wall is *not* engaged in a goal-directed activity? Assiduity and the ability to recognize that the end-state has been achieved count more heavily than insight or brilliance in

execution. We do ‘expect’ intelligent animals to improve and prune their repertoire, but that is because we expect intelligent animals to be learners as well as goal-havers.

It is possible then to make a clear distinction between genuine goal-

directed behaviour on the one hand and goal-terminated behaviour coupled with the capacity to learn on the other? The vague notion of goal-direction suggests that having a goal is also having rationales for the means one attempts, having the ability to distinguish appropriate from inappropriate courses of action, but this is a misleading suggestion, for one need not have a particular goal in order to be able to decide what *would be*

appropriate courses of action if one did have the goal. The ratiocinative capacity is separable from the having of goals, at least in man, and surely

we want to maintain a distinction between well-reasoned goal-directed activity and ill-reasoned or unreasoned goal-directed activity.

But at least, it may be argued, one must know what one’s goal is before one can even begin to bring into play the separate capacity of

ratiocination, while in goal-terminated activity one can be entirely in the dark about what the end-state is until it is reached. This seems like a

promising mark of difference until it is asked what the criteria are to be for knowing one's goal. For people we generally want to say that being

able to state that *p* is a necessary condition for knowing that *p* (the one trivial counter-example being the paralysed aphasic who has no means of communication). If this is held to be a necessary condition for knowing, then only human beings can know, and hence only human beings can be

goal-directed. (Since there is a great difference between being

incapacitated by aphasia and being, like a dumb animal, a constitutional non-speaker, animal knowledge cannot be brought in under the

counterexample mentioned above.) Now perhaps this is what we want.

Perhaps we want to say that only human beings exhibit true goal-directed behaviour, and the capacity for goal-direction, like the capacity for

erudition or eloquence, is reserved for language users only. This is

unconvincing. Nothing in our vague notion of goal-direction suggests that the use of language is a prerequisite, and we do not, I should think, want to rule out as goal-directed the more remarkable activities of higher animals and very young children. If so, then we must find a different sense of

knowledge, one that does not require that what is known be statable by the knower. We must find some other behavioural criteria for knowing, and,

more particularly, behavioural criteria for knowing one's goal.

What other behavioural cues for knowing one's goal could there be but

taking steps towards achieving the goal and stopping when the goal is

achieved? If these are the only relevant cues we are back to intelligent goal-terminated activity, for the behavioural evidence will be the same for both it and putatively 'genuine' goal-directed activity. Deciding whether a particular animal is exhibiting goal-directed behaviour will hinge on how

we interpret its motions: are they sufficiently directed *towards* achieving the goal? How we answer this question depends on our evaluation of the

appropriateness of the attempts, e.g., does the animal succeed in

recognizing partial progress towards its goal, does it learn, does it abandon too early what to us seem like promising avenues? Answers to these

questions admit of degrees and disagreement. If we set the standards high, only well-reasoned goal-directed activities will count as goal-directed at all, and once again the distinction between well-reasoned and ill-reasoned goal-directed activities is lost. If we do not set standards but allow the notion to admit of degrees, then we are left saying that some activities are more truly goal-directed than others, an unhappy way of looking at things; either one has a goal or not. An alternative and better way of describing such activities would be to describe some goal- *terminated* activities as more appropriately marshalling their sub-routines than others.

It is tentatively proposed, then, that the GPS model of goal-direction

can do justice to the observed behaviour of animals. Nothing an animal

could do short of giving us a disquisition on its goals and methods would give us evidence pointing to more marvellous control systems than those sketched for GPS. I make the proposal tentative not because I intend to replace it later, but because it is difficult to see what, if any, limits can be set on the heuristics and learning potentials of GPS-styled systems, and so it is difficult to say with any certainty that such systems would or would not be adequate to model any animal behaviour yet to be discovered. The behavioural evidence so far culled by the psychologists does not suggest that animals are capable of behaviour superior in shrewdness or different in style from the behaviour of such a control system. The question of

whether human beings, with their greater sophistication, require different sorts of control systems is a very difficult question, but some progress will be made on it indirectly in Chapters 7, 9 and 10.

Supposing, then, that we accept GPS as an adequate simple model of

goal-direction, the next question is whether a control system having the general characteristics of GPS could be produced by the evolutionary

system proposed. The first problem is the specification of an end-state. If animals were in the habit of scrambling about until they were presented by a unique retinal projection or olfactory stimulus, or any other simple and easily described peripheral afferent, the problem would be simple. Such a stimulus could serve as the terminating stimulus for whatever efferents had been operating, and its occurrence could be designated the end-state.

But animals do not do this. Their goals are not often the experiencing of particular peripheral sensations, with the possible exception of hunger satiation signals. Animals are more apt to have the goal of finding their way home, or to the feedbox, or of reaching safe ground, or of breaking the

clam-shell, and no unique sensory presentation or even finite disjunction of such presentations would serve to signal the achieving of these goals.

For these relatively sophisticated goals there is no hope of finding a

peripheral neural state that would serve as a suitable end-state. So the peripheral stimulation must be processed into something more

sophisticated.

This is the problem of pattern recognition and stimulus generalization, and considerable progress has been made in exhibiting the power of

stimulated neural nets to perform these tasks. The details need not concern us, for all that is important to us is that the output activity of the afferent nervous system (i.e., the central activity that becomes the *input* to the efferent nervous system) should be capable of being determined not by

specific peripheral patterns but by external conditions described more

generally. The successes of relatively small pattern analysing devices built of modules analogous to, but much simpler than, neurons indicate that the immense and highly complex neural net that makes up the afferent

nervous system is fully equal to this task. Supposing the afferent portion of the brain of a higher animal to be such a neural net, it should be capable of producing output states sufficiently interpreted relative to peripheral stimulation to serve as specifications of goal-states. For example, while particular neural signals near the retina might fire normally if and only if vertical lines predominated on the retinal projection, signals at a higher level in the neural net would fire normally if and only if the animal were surrounded by large vertical objects, and at a higher level still there would be signals – that is, as we have seen, concatenations of severally

ambiguous signals – firing normally if and only if the animal were safely hidden among the white pines. Such high-level activity or some state

resulting from it could serve as an end-state for goal-directed behaviour, e.g., running until safe ground is reached.

The question that remains is whether such systems of subroutines and

end-states could evolve within the individual brain under the principles we have already established. A detailed account of mechanisms and structures that would be required for such ‘learning’ would take us much farther into the area of detailed empirical speculation than I wish to go. Our ‘model’

of the evolving brain would have to be made much more detailed and even fitted out with variables to which numerical weights could be assigned in formulae, and so forth. Since it is very much in our interests to keep all

such hypotheses as general as possible, such specifications would work to defeat our purposes. In a very general way, however, we can see what

direction such an evolution might take.

It is a common belief among psychologists that the normal behaviour of

animals and perhaps even of man is divisible into hierarchies of patterns generated from the animal’s basic needs – essentially food, defence and procreation. This belief can be found in various forms in philosophical and

psychological theories dating back at least to Aristotle, and is, of course, the offspring of a more homespun belief that the purposes of man and

beast are nested in a few or perhaps just one basic purpose. One saws the plank to build the door to put on the house to keep it secure to protect one's health to stay alive. Many models have been proposed to account for this characteristic of behaviour, and typically they have been hydraulic in inspiration, with pressures being channelled this way and that. Such

hydraulic models, with their mysterious fluids – humours, libido, *élan vital* – being shunted about, are, of course, very much out of date today, but they do suggest a general principle of generation that might find many different embodiments within the more powerful and versatile framework

of information theory.

A nesting of hierarchies could probably be generated from a single

'seed', a pre-wired set of controls for some one simple pattern of goal-directed behaviour. For example, if some 'hunger'-afferent had the

genetically established effect of producing continuing, widespread, but

'unspecified' efferent activity – leading, one might suppose, to behaviour in the form of random scrambling about – this efferent activity could be pruned along the lines of various different ways of getting food by the evolutionary principles we have already proposed. Once these different

sub-routines had been established, sub-goals within them could develop

from the convergence of new afferent and efferent activity, and so forth, spawning in pyramidal fashion a series of goal-direction controls in

overall response to the obstacles in the environment preventing the direct achievement of the pre-wired end-states. In such a view the *pressure* of the hydraulic models still exists in not entirely metaphorical form: the

afferent initiation of the behaviour produces a pressure of efferent activity which seeks to be relieved via various routes, these routes controlling

various attempts at achieving the goal.

I do not want to lean at all heavily on this rather cloudy view of goal-generation, but only use it as an illustration of one possible avenue for the researchers and programme devisers. In all likelihood considerably *more* hierarchical structure is genetically transmitted; for example, food-seeking tropisms and reflexes (including the sucking reflex of a baby) of considerable complexity are very widespread in nature, and in higher

animals these could serve as a much more elaborate starting point for

goal-generation than was suggested above.

The preliminary sketch of a centralist theory of behaviour developed in this chapter is intended only to reveal the general shape such a theory must take if it is to deal with the problems set for it by what are largely *a priori* conditions. Although there has been some guarded comment about the

'significance' of certain neural impulses, and the 'ambiguity' of certain others, no strict justification has been yet proposed for what must be the crux of any centralist theory: the ascription of content or meaning to

particular central states of the brain.

4

THE ASCRIPTION OF CONTENT

IX FUNCTION AND CONTENT

So far the argument has been that if there is to be a *rapprochement* between the extensional physical sciences and 'the language of the mind'

– whether our ordinary Intentional discourse or the statements of a

'science of Intention' – we must find a rationale and justification for ascribing content to certain internal states and events of the behavioural control system. And since Intentional explanations *presuppose* the appropriateness of the sequences of events they purport to explain (see

Chapter 2), part of the burden of such content ascription is providing an account of the generation of structures to direct these generally

appropriate sequences. It was to meet this requirement that we proposed hypotheses about evolution, both of species and of neural structures. Put another way, since environmental significance is extrinsic to any physical features of neural events, and since the useful brain must discriminate its events along lines of environmental significance, the brain's

discriminations cannot be a function of any extensional, physical

descriptions of stimulation and past locomotion alone. Rather, some

capacity must be found in the brain to generate and preserve fortuitously appropriate structures. It was then argued that a close analogue of natural selection of species would be a system that could provide this capacity and could itself be provided for by natural selection of species. The system was developed just enough to provide some answer to the question of

whether it could control goal-directed behaviour, but it will provide us with footholds for the next task: determining the conditions under which one could justifiably ascribe content to neural states.

For a start it is clear that for any system to be called Intentional it must be capable of discriminating and reacting to fairly complex features of its environment (e.g., external physical objects and not just changing

conditions – temperature, contact, pressure – on its outer surface), and for any system to do this it must be capable of interpreting its peripheral stimulation. That is, it must be capable of producing within itself states or events that normally co-occur with generalized conditions of objects

within the system's perceptual field. I do not think this is a *formal* requirement for any Intentional system so much as one designed to satisfy our intuitions; no system that lacked this capacity could engage its

environment in ways interesting and sophisticated enough to make it *plausible* to say that it had beliefs, desires, intentions – even if in the end we

could find no logically necessary trait for, say, belief, that the system lacked. This capacity for afferent analysis does not suffice in itself, however, to establish a system as Intentional, for the information produced by such an analysis, for all its abstraction from its source in peripheral stimulation, will still be only non-intelligently held information unless something else is added. The something else is a certain association

between the results of afferent analysis and structures on the efferent side of the brain. This can be brought out by example. Suppose that in an

organism O there is a particular highly interpreted afferent output A

(summing, we can suppose, signals from visual, tactile and olfactory

sources) that fired normally if and only if food was present in O's

perceptual field. The firing of A might have any of a vast number of

effects on O's behaviour. If it happened, for example, to have the effect of terminating a series of 'seeking' sub-routines and initiating a series of other, 'eating' sub-routines, we would have evidence for saying that O had achieved its goal of finding food, had recognized that the goal was

achieved, had discriminated the presence of food *as* the presence of food.

If, on the other hand, A did not have this effect, if O did not commence eating or in other ways behave appropriately to the presence of food under the circumstances, then regardless of any evidence we might have about

the specificity of the stimulus conditions determining the firing of A, there would be no reason to say that the animal had discriminated the presence of food *as* the presence of food.

This point has often been missed, probably because of a misplaced

analogy between our introspective experiences while problem-solving and the state of affairs that exists in the brain. The point is missed when some S-R behaviourists pose their most intractable problem: how does the novel stimulus (meaning what it does) *get to* or *select* the appropriate response?

This is a hopeless question, for it presupposes an impossible state of affairs in the brain, in which the brain somehow recognizes or discriminates a stimulus as, say, one of pain or of a white triangle or of the dead-end of a maze alley, but does not 'know' yet what the appropriate motion is in the face of such a stimulus. This view of the situation is apparently harmonious with our experience in that we often come upon things in our environment about which we wonder 'what shall I do about that?', so it seems plausible enough that discrimination of stimuli and doing something about them are perfectly separable. But this extrapolation from human experience is not justifiable on the level of explanation involved when one is talking about brains rather than people. In the brain, discrimination of afferents according to their significance just *is* the production of efferent effects in differential response to afferents, and hence it does not make sense to suppose that prior to the production of an efferent event or structure the brain has discriminated its afferents *as* anything at all.

No afferent can be said to have the significance 'A' until it is 'taken' to have the significance 'A' by the efferent side of the brain, which means, unmetaphorically, until the efferent side of the brain has produced a response (or laid down response controls) the unimpeded function of which would be appropriate to having been stimulated by an A. This is *not* the epistemological point that as behaviourists *we cannot tell* whether the organism's brain has discriminated its stimulus as having the significance 'A' until the organism manifests this in its behaviour, but the logical or conceptual point that *it makes no sense* to suppose that the discrimination of stimuli *by their significance* can occur solely on the afferent side of the brain. The epistemological point is a canon of the experimental method in psychology: since the animal cannot tell us whether it can tell a circle from a square, we must set up the situation so that its behaviour tells us.

This canon, as it stands, hides an ambiguity. Surely there is another

alternative which we are prevented from using only by the limits on our present research techniques. We could in principle record the afferent

activity in the animal when its eyes were presented with circles and

squares and, on the basis of vast knowledge of the principles of afferent function, determine that the animal's afferent analysis system had unique and different outputs for circles and squares. Would this show that the animal discriminated circles from squares? In one sense, it would. This is the sense of discrimination of interest in research into pattern recognition devices, where all that is at issue is whether or not the system is capable of producing outputs – whatever they may be – that co-occur with the critical patterns of the inputs. In principle we could know that in this sense an animal could discriminate circles from squares without ever examining its overt behaviour. This is not yet discrimination by significance, however.

We would not give as the conclusion of this experiment that the animal

could discriminate circles *as* circles and squares *as* squares. Furthermore, for all animals lower than human beings there is no behavioural

experiment we could perform that would have this as its conclusion, since circles and squares, even under laboratory conditions, could have no

bearing *as* circles and squares on the life and activities of the animal. They could have bearing as left-turn indicators or as warnings of an electric shock, but not as circles and squares. This can be seen by contrasting

circles and squares with food pellets. One could set up an experiment in which a food pellet served as a left-turn indicator for a rat in a maze, and once the rat had learned this we could say its behaviour showed that it discriminated the food pellet as a left-turn indicator. To discriminate the food pellet as food, on the other hand, is to try to eat it. There is something appropriate a rat can do with a food pellet such that it makes a difference whether it is a food pellet or a marble, but there is nothing a rat could do with a circle such that it makes a difference whether it is a circle or a square

or a triangle. This limitation is due, of course, to the very limited interests and activities of rats. Were rats interested in making wagon

wheels the situation would be different. The significance an item in the environment can have to a creature is limited by the creature's behavioural repertoire, but this limitation only comes into force at the level of

afferent-efferent intermeshing, which is, therefore, the first point at which we can speak of discrimination by significance. Since, then, effects on behavioural controls are *conceptually* required for there to be

discrimination by significance, and since a stimulus, as a physical event, can have no intrinsic significance but only what accrues to it in virtue of the brain's discrimination, the problem-ridden picture of a stimulus being *recognized* by an animal, meaning something *to* the animal, prior to the animal's determining what to do about the stimulus, is a conceptual

mistake.

An idealized picture of content ascription emerges from this from which we can draw some conclusions before complicating it out of existence.

The content, if any, of a neural state, event or structure depends on two factors: its normal source in stimulation, and whatever *appropriate* further efferent effects it has; and to determine these factors one must make an assessment that goes beyond an extensional description of stimulation and response locomotion. The point of the first factor in content ascription, dependence on stimulus conditions, is this: unless an event is somehow

related to external conditions and their effect on the sense organs, there will be no grounds for giving it any particular *reference* to objects in the world. At low enough levels of afferent activity the question of reference is answered easily enough: an event refers to (or reports on) those stimulus conditions that cause it to occur. Thus the investigators working with

fibres in the optic nerves of frogs and cats are able to report that particular neurons serve to report convexity, moving edges, or small, dark, moving objects because these neurons fire normally only if there is such a pattern

on the retina. 1 However mediated the link between receptor organ and higher events becomes, this link cannot be broken entirely, or reference is lost.

The point about the link with efferent activity and eventually with

behaviour is this: what an event or state ‘means to’ an organism also

depends on what it *does* with the event or state. This suggests a tempting but not altogether reliable analogy with the logical distinction of extension and intension: the stimulus conditions dependence ensures that neural

‘expressions’ will have reference or extension, while the efferent effect dependence ensures that they will have sense or intension. Our paradigm here is the case of the simple nervous system of strain A, in which there is an inherited arc linking a certain stimulus with a withdrawal motion, and the stimulus conditions of which are, as a general rule, harmful to the well-being of strain A. The effect of this link and the conditions under which it operates give us reason for calling the afferent side of the arc a signal of pain, or perhaps danger, but there would be no reason for this ascription if the organism responded inappropriately or not at all to the stimulus, whatever the conditions of stimulation. That the stimulus did not mean danger to *him* would be abundantly clear from his reaction. The criterion for intelligent information processing must involve this

behavioural link – however mediated – since propitiousness or

adaptiveness of behaviour is at least a necessary condition of intelligence.

This immediately establishes a limit on the events and states within the brain to which the investigator can ascribe content. Where events and

states appear inappropriately linked one cannot assign content at all, and so it is possible that a great many events and states have no content,

regardless of the eventual effect they have on the later development of the brain and behaviour.

This point is important enough to be worth further development. Let us concoct an artificial case in which the behaviour is wildly inappropriate to the perceptual environment. Fido, who has not been fed all day, is handed a large chunk of beefsteak, but instead of eating it he carefully gathers together a little pile of straw, puts the meat in the middle, and sits down on the meat. Now suppose further that we have voluminous data on Fido's

neural states. Afferent state A is the outcome of the convergence of

olfactory, visual and tactile stimulation, and is the normal outcome of afferent analysis when Fido in the past has discriminated food. But this time its efferent continuation leads to the bizarre behaviour. Since Fido has not behaved appropriately, we cannot say that state A has the content (roughly) 'this is food' for him, but if not, no other candidate is supported either. Fido's behaviour would be appropriate to a belief that the beef was an egg and Fido was a hen, and since state A has the efferent effect

governing this behaviour it might seem that solely on the basis of our

second factor, the generation of behaviour controls, we can ascribe content to state A: 'this is an egg and you are a hen'. But Fido's behaviour is also appropriate to other beliefs, e.g., 'this is beef, but if you pretend it's an egg you'll get twice as much beef tomorrow', or 'it is worth starving to throw these psychologists into confusion' or 'sitting on beef improves the

flavour'. Since any behaviour will be appropriate to a variety of different beliefs and desires, the only feature that can be counted on to determine the correct hypothesis will be the afferent source of the structure that governs the behaviour, and the afferent source will favour one of the

hypotheses only in the event that the behaviour is appropriate to the

conditions of this source. Where there is an inappropriate liaison, the response to the environment 'makes no sense', and, since it makes no

sense, no Intentional (putatively sense-making) account of the liaison will be justified.

So, one can only ascribe content to a neural event, state or structure

when it is a link in a demonstrably appropriate chain between the afferent and the efferent. The content one ascribes to an event, state or structure is not, then, an extra feature that one *discovers* in it, a feature which, along with its other, extensionally characterized features, allows one to make predictions. Rather, the relation between Intentional descriptions of

events, states or structures (as signals that carry certain messages or memory traces with certain contents) and extensional descriptions of them

is one of *further interpretation*. If we relegate vitalist and interactionist hypotheses to the limbo of last, desperate resorts, and proceed on the

assumption that human and animal behavioural control systems are only

very complicated denizens of the physical universe, it follows that the events within them, characterized extensionally in the terms of physics or physiology, should be susceptible to explanation and prediction without any recourse to content, meaning, or Intentionality. There should be

possible some scientific story about synapses, electrical potentials and so forth that would explain, describe and predict all that goes on in the

nervous system. If we had such a story we would have in one sense an

extensional theory of behaviour, for all the *motions* (extensionally characterized) of the animal caused by the activity of the nervous system would be explicable and predictable in these extensional terms, but one thing such a story would say nothing about was *what the animal was*

doing. This latter story can only be told in Intentional terms, but it is not a story about features of the world *in addition to* the features of the extensional story; it just describes what happens in a different way.

Supposing one could have complete knowledge of the mechanics of a

computer without the slightest inkling of the rationale of its construction, one would be in a similar situation: one would see type pressed against

paper, relays open and close, and all this would be predictable and

explicable in terms of physics, but one would have nothing to say in this account about the logic of the operations, about adding, subtracting and comparing, or even about *operations* at all.

A solely biological, non-Intentional theory of behaviour should be

possible in principle, but it would be mute on the topic of the actions (as opposed to motions), intentions, beliefs and desires of its subjects.

Moreover, the theory would be very difficult to get to without the

understanding provided by the Intentional ascriptions of content. Thus one motive for centralism is that it can provide the physiologists with an

invaluable heuristic advantage, as the physiologists have been quick to see; if they cannot view neural events as *signals* or *reports* or *messages*, they are left with almost no view of brain function at all. Were the

physiologist to ban all Intentional coloration from his account of brain functioning, his story *at best* would have the form: functional structure A has the function of stimulating functional structure B whenever it is

stimulated by either C or D … No amount of this sort of story will ever

answer questions like why rat A is afraid of rat B, or how rat A knows which way to go for his food. If one does ascribe content to events, the system of ascription in no way interferes with whatever physical theory of function one has at the extensional level, and in this respect endowing events with content is like giving an interpretation to a formal

mathematical calculus or axiom system, a move which does not affect its functions or implications but may improve intuitive understanding of the system. 2

The heuristic value of giving an Intentional interpretation to events

varies, of course, with the complexity of the events and their remoteness from the periphery of the nervous system. There is nothing to be gained by assigning content to the last-rank motor impulses that stimulate muscle contraction, for example. Giving such an event the imperative message

'contract now, muscle!' does little to clarify what is going on. Deeper in the brain, however, characterizing a state or event or structure not only as a physical entity operating under certain causal conditions but also as, for example, a specification of a goal or description of the environment or order to perform a certain task would be virtually the only way of 'making sense' of neural organization. More important to us here, however, than any aid and comfort Intentional interpretations may give the investigator is the matter of principle. If the idea of content ascription is sound in principle, regardless of how messy or useless it is in practice, it allows the conclusion that natural physical organisms are, with no help from

Cartesian ghosts or interacting vital forces, Intentional systems.

The ideal picture, then, is of content being ascribed to structures, events and states in the brain on the basis of a determination of origins in

stimulation and eventual appropriate behavioural effects, such ascriptions being essentially a heuristic overlay on the extensional theory rather than intervening variables of the theory. A centralist theory would consist of two levels of explanation: the extensional account of the interaction of functional structures, and an Intentional characterization of these

structures, the events occurring within them, and states of the system

resulting from these. The implicit link between each bit of Intentional interpretation and its extensional foundation is a hypothesis or series of hypotheses describing the evolutionary source of the fortuitously

propitious arrangement in virtue of which the system's operation in this instance makes sense. These hypotheses are required in principle to

account for the appropriateness which is presupposed by the Intentional interpretation, but which requires a genealogy from the standpoint of the

extensional, physical theory.

This ideal picture will provide a basis for discussion in subsequent

chapters, but first there are complications to it which must be described since they have important implications of their own. First, the problem of tracing the link between stimulus conditions and internal events far from the periphery should not be underestimated. Even discounting the

'ambiguity' which was seen in Chapter 3 to infect neural signals generally, it is not to be expected that central events can be easily individuated in such a way that they have unique or practically unique sources in external stimulation. Suppose we tentatively identify a certain central event-type in a higher animal (or human being) as a perceptual report with the content

'danger to the left'. Now probably in higher animals and certainly in

human beings we would expect the *idea* of 'danger to the left' to be capable of occurring in many contexts: not only in a perceptual report, but also as part of a dream, in hypothetical reasoning ('what if there were danger to the left'), as a premonition, in making up a story, and of course in negative form: 'there is no danger to the left'. What is to be the

relationship between these different ways in which this content and its variations can occur? Are we to hope for one extensionally characterized event-type an instance of which occurs whenever this idea in any of its guises occurs, or will the different contexts correlate with regular, law-governed variations of our initial event-type, or will there be one event-type, presumably the original perceptual report event-type, which

systematically spawns the second-order event-types which are the signals of imagination, reasoning and so forth? What of belief that there is danger to the left? Belief is not an event, something that happens, but a state (which can sometimes be *dated*, but cannot be swift or slow), so are we to suppose that the *state* with this belief-content is established in any typical or regular way by our perceptual report event-type? Certainly for any

event or state to be ascribed a content having anything to do with danger to the left, it must be related in some mediated way to a relevant stimulus source, but the hope of deciphering this relation is surely as dim as can be. 3

The problem with behavioural effects is similar. I have held that the

claim to *intelligent* use of information depends on there being appropriate

continuations or effects of signals, but how appropriate must they be, and how direct or indirect? How are we to measure potential effects on

behaviour without a total knowledge of the functioning of the nervous

system? Certainly the less direct the afferent-efferent links are the more difficult it will be to discover that they are at all appropriate. The more room there is for mediation and complexity, the more potentially

intelligent a creature will be, but also the more difficult it will be to find detailed evidence that this intelligence in particular cases is due to this or that feature of its neural organization.

An event, state or structure can be considered to have content only

within a system as a whole, and it is this fact that virtually precludes the possibility of content ascription to events, states or structures that are relatively central in any large nervous system. Until one has traced their normal causes and effects all the way to both the afferent and efferent peripheries, one can have no inkling at all of their content. Near the

peripheries one can ignore one condition or the other and so determine

content of neuronal activity to a first approximation, as e.g., reporting a dark object in the visual field or ordering the raising of a leg, but by ignoring the eventual effects of the former and the central causes of the latter one leaves untouched the fundamental problem of how the brain *uses* information intelligently, and so one cannot be said to have determined the meaning of the event within the system as a whole.

The task of ascribing content can be divided into two parts: the

individuation by function of neural structures, events and states, and the subsequent framing of messages or contents for them. We have seen that a number of problems make the first half of the task all but impossible. For one thing, the relevant functions that must be determined are not local but global, extending to the peripheries. For another, the events and states that would be good candidates for content-bearers are, at least in the central areas, compound, ambiguous and apparently continuously changing.

Difficulties of a different sort affect the second half of the task.

X LANGUAGE AND CONTENT

Assigning content to an event must be relating the event to a *particular* verbal expression. This could be done somewhat fancifully by using the

form of direct quotation. The signal says, or tells the brain, ‘food straight ahead’ or ‘turn to the left’ or ‘there’s a pain in your left foot’. Only apparently more austere would be assignments in terms of indirect

quotation or propositional attitude: a signal is to the effect that …, or reports that …, or commands that …, and these are Intentional contexts, as are the forms: reports the presence of x, commands the x to y, etc. This is the point of centralism, to relate meanings to events, and this involves *expressing* the content of events, since content cannot be *described*. But then which expressions shall we use?

At what level of afferent stimulus analysis in the neural net, for

example, shall we move from content in terms of events in the sense

organs to content in terms of events and objects in the external world?

When do signals report not just patterns of excitation on the retina but things seen? In the case of the frog, for example, when do we say the

analysis of stimulation has produced a signal about a moving dark object in the environment (the fly) rather than a moving dark area on the retina?

It might seem that the answer is that object reference is permissible after convergence of signals from both eyes, or from several sense organs, but the frog will commit itself to a behavioural response on the basis of

information from one eye alone. Here our semantic analogy to the effect that reference is determined by stimulus conditions and sense by efferent continuations breaks down. Here the shift from a retinal reference to an object reference must depend on what effect a signal has on behaviour. It is tempting in these cases to confuse a psychological question with an

epistemological question. Must we lift, taste, smell and hear an object in addition to seeing it before we have ‘conclusive evidence’ that it is a concrete object in the world, or is seeing enough? Fortunately we do not require conclusive evidence of objectification, whatever that might be, before we act, or we would all starve to death. What our senses ‘tell’ us is not what they *prove* to us, and the question facing the centralist is what the organism ‘takes the signal to mean’.

Even if there is a comfortable way of deciding when to raise the level of information to objective reference, there remains the question of how to describe the objects referred to. Let us consider another necessarily crude hypothetical example. A centralist of the future has access to the neural events in Fido’s brain and observes him refusing to venture out on to thin ice to retrieve a succulent steak. He has the following information: an afferent event of type A, previously associated with visual presentations of steaks, has continuations which trigger salivation and also activate a

control system normally operating when Fido is about to approach or

attack something, but this efferent continuation is inhibited by signals with a source traceable to a previous experience when he fell through thin ice. That is, the centralist has information regarding neural functioning that puts him in a strong position to say that Fido’s behaviour is

determined in this case by the stored information that it is dangerous to walk out on thin ice. Such an account would be better substantiated than, for example, ‘Fido did not notice the steak’, ‘Fido has an aversion to

smooth horizontal planes', 'Fido is overcome by Weltschmerz'. On the

basis of his vast knowledge of the functional interrelations in Fido's

nervous system, the centralist assigns certain contents to certain events and structures. Roughly, one afferent signal means 'there's a steak', its continuation means 'get the steak', some structure or state stores 'thin ice is dangerous' and produces, when operated on by a signal meaning 'this is thin ice', another signal meaning 'stop; do not walk on the ice'. (The point about stimulus conditions and behavioural effects determining content

comes out particularly clearly here; no structure or state could be endowed with the storage content 'thin ice is dangerous', no matter how it had been produced, if the input of 'this is thin ice' did not cause it to produce an appropriate continuation, such as 'do not walk on the ice'. In the absence of such appropriate functioning one would be bound to conclude that the animal had failed to remember his previous experience, had failed to store intelligently that information, even if there were some clearly identifiable trace in the brain owing its origin to the earlier experience.)

As soon as we consider any standards of accuracy in content ascription, the particular choices of the centralist in this example begin to look too crude. Does Fido really discriminate the object *as* a steak, or would 'meat'

or 'food' have been more accurate choices? Presumably the signal's

stimulus conditions are more specific than would be implied by the word

'food', and we can expect the dog to show more interest in steak than in dog biscuits, so 'food' does not seem to be a good choice from the point of view of either stimulus conditions or behaviour, but 'meat' suggests too much. Surely the dog does not recognize the object as a butchered animal part, which is what the word 'meat' connotes, and 'steak' has even more specific implications. Should we be worried by these implications? Yes, if what we are trying to do is 'specify the concepts' that operate in the dog's direction of behaviour. What the dog recognizes this object as is

something for which there is no English word, which should not surprise us – why should the differentiations of a dog's brain match the

differentiations of dictionary English?

It might seem that we could get at the precise content of the signal by starting with an overly general term, such as 'food', and adding

qualifications to it until it matches the dog's differentiations, but this would still impart sophistications to the description that do not belong to the dog. Does the dog have the concept of nourishment that is involved in the concept of food? What could the dog do that would indicate this?

Wanting to get and eat *x* is to be distinguished from recognizing *x* as food.

These hair-splitting objections might lead the zealously rigorous centralist to formulate artificial languages for expressing the content of the events and states he isolates, but to go to such efforts in the name of precision is to lose sight of the essential point and burden of centralism.

The centralist is trying to relate certain Intentional explanations and descriptions with certain extensional explanations and descriptions, and the Intentional explanations that stand in need of this backing are nothing more than the rather imprecise opinions we express in ordinary language, in this case the opinion that Fido's desire for the steak is thwarted by his fear of the thin ice. If the centralist can say, roughly, that some feature of the dog's cerebral activity accounts for his desire to get the steak, and some other feature accounts for his fear (inculcated by certain past

experiences) of what he takes to be thin ice, he will be matching

imprecision for imprecision, which is the best that can be hoped for.

Precision would be a desideratum if it allowed safe inferences to be

drawn from particular ascriptions of content to subsequent ascriptions of content and eventual behaviour, but in fact no such inferences at all can be

drawn from a particular ascription. Since content is to be determined in part by the effects that are spawned by the event or state, the Intentional

interpretation of the extensional description of an event or state cannot be used by itself as an engine of discovery to predict results not already discovered or predicted by the extensional theory. Ascriptions of content always presuppose specific predictions in the extensional account, and

hence the Intentional level of explanation can itself have no predictive capacity. That is, while it is true that *if* a person believes that A is the only way to get B, and *if* he wants B, it follows or can be predicted that he wants A, the centralist cannot use such an Intentional prediction to predict further events in the nervous system, for he could have no evidence that the antecedents of the hypothetical were true (and precise) unless he had already determined or predicted (via his extensional theory) the existence of the state which he would associate with wanting A, and so forth all the way to behavioural manifestations. Since Intentional explanations

presuppose appropriateness or rationality, rational coherence is a logical requirement of content ascriptions, but it is no logical requirement of neural function (which may suffer breakdowns or be infelicitously

organized in the first place), and therefore inferences made at the

Intentional level will be borne out only when neural functional

organization achieves ‘ideal’ rationality, something for which there is no guarantee, and no way to check independently of extensional level

determinations of function. From any portion of the Intentional story a further portion can be generated only on the assumption that the

ascriptions of content so far made are ‘accurate’, and to test this

assumption one must see if what one generates on the basis of these

ascriptions is borne out by details of the extensional story. The ascription of content is thus always an *ex post facto* step, and the traffic between the

extensional and Intentional levels of explanation is all in one direction.

This feature can be easily overlooked by investigators in memory

mechanisms, who occasionally speak as if they were looking for word-

analogues and sentence-analogues in the brain. A sentence token (a

particular occurrence of a sentence) is a token of a particular sentence type in virtue of its having certain syntactic parts (word tokens) and a certain syntactic structure (the ordering of the word tokens), and these features of the thing or event, the sentence token, serve to restrict and determine – in ways very difficult to describe – the function the thing or event has within a particular system, say Jones and Smith conversing in English. Thus when Jones says ‘pass the salt’, the likely effect of this utterance event on the

system Jones-Smith is in part based on internal (in this case phonological) traits of the event to which we ascribe content. The event has syntactic parts that can be read off (by anyone who understands English). There is no guarantee, however, that the things and events making up the

Intentional system that is a particular creature will have analogous

syntactic parts or structures at all and, if they do not, there is no guarantee that they will have their functions restricted in ways much like the ways in which sentence tokens have their functions restricted. It is possible,

perhaps, that the brain has developed storage and transmission methods

involving syntactically analysable events or structures, so that, for

example, some patterns of molecules or impulses could be brain-word

tokens, but even if there were some such ‘language’ or ‘code’ or what

Zeman calls ‘the brain writing which people have in common regardless of their nationality and other differences’, [4 t](#)here would also have to be mechanisms for ‘reading’ and ‘understanding’ this language. Without such

mechanisms, the storage and transmission of sentence-like things in the brain would be as futile as saying ‘giddyap’ to an automobile. These

reading mechanisms, in turn, would have to be information processing

systems, and what are we to say of *their* internal states and events? Do *they* have syntactically analysable parts? The regress must end eventually with some systems which store, transmit and process information in non-syntactic form. Of all the common analogies used to describe the brain, the analogy of a community of correspondents (which is the inevitable

suggestion whenever there is talk of codes and languages in the brain) is the most far-fetched and least useful. It has the disadvantage of merely postponing the central problem before us by positing unanalysed

mananalogues as systematic elements in that which we are trying to

analyse, namely Man. The ‘little man in the brain’, Ryle’s ‘ghost in the machine’, is a notorious non-solution to the problems of mind, and

although it is not entirely out of the question that the ‘brain-writing’

analogy will have some useful application, it does appear merely to

replace the little man in the brain with a committee. 5

The ‘brain-writing’ view obscures the important truth that the capacity of a language to store and transmit information (in books in libraries, in speeches and on signboards) is dependent upon the existence of non-linguistic means of storing and transmitting information. Information is not preserved in a sentence like a fossil in a rock; a sentence is a vehicle

for information only in that it is part of a system that necessarily includes sub-systems that process, store and transmit information non-linguistically. Whether these sub-systems are whole men or whole nervous systems or certain of their parts is an empirical question, but there can be no doubt that there are such sub-systems. Within such sub-systems the

association of verbal messages or contents with events and states can be given a rationale only by pointing out the effective contribution of these events and states to the direction of behaviour that is ultimately

appropriate to the survival of the system as an organism in the world.

This fact puts the Intentionalist thesis of irreducibility in a new light.

Initially the question of whether Intentional accounts of behaviour and

'mental' events could be reduced to or paraphrased into extensional

accounts was seen as the question of whether events and states of the

nervous system could be assigned meanings or ascribed contents, and

assigning meanings was seen as associating events or states with verbal expressions. Verbal expressions, however, are not the ultimate vehicles of meaning, for they have meaning only in so far as they are the ploys of

ultimately non-linguistic systems. The inability to find precisely worded *messages* for neural vehicles to carry is thus merely an inability to map the fundamental on to the derived, and as such should not upset us. Although this examination of the problem of the ascription of content has yielded the conclusion that no rigorous, predictive way of ascribing content is possible, a rationale for a looser but still explanatory assignment of

meanings to events and states has been developed. This is enough to blunt the point of the Intentionalist thesis. If in a sense the thesis still stands, it no longer should have the effect of suggesting an unbridgeable gap

between the mental and the physical – whether this is construed as a

radical dualism of phenomena, or of sciences, or of modes of description or explanation. For although no neat synonymy or correlation between

Intentional and non-Intentional sentences has been discovered or

proposed, sense has been made of the lesser claim that certain types of physical entities are systems such that their operations are *naturally* to be described in the Intentional mode – and this, only in virtue ultimately of their physical organization. The force of 'naturally' here is this: although such systems are ultimately amenable to an extensional theory of their

operations, their outward manifestations are such that they can be

intelligibly described at this time, within our present conceptual scheme, only in the Intentional mode.

The Intentional mode, along with the extensional mode, is a given in our conceptual scheme, and as such it must serve as both a starting point and an at least *pro tempore* reference point for explanations that go deeper than our ordinary remarks about behaviour or minds. The role of the

Intentional mode as a given can perhaps best be understood by looking

back to earlier times when its scope as a given was wider, when our

animistic ancestors spoke Intentionally about rivers, clouds, fires,

mountains. From our present vantage point it is easy enough to say that talk of the river's desire to reach the sea was a clear overextension of the mode, but the fact remains that there was a time when this was a

phenomenon, *Intentionally characterized*, to be explained – or explained away. And until we stopped speaking seriously in this way about rivers, the Intentional characterization remained a reference point for explanations; what had to be explained was the river's desire. From our present vantage point it would make no sense to say that the Intentional mode applied

today to people, animals and occasionally to computers is similarly an

over extension of the mode, for the 'correct' scope of the Intentional mode is determined at any time by the current conceptual scheme. Intentionally characterized phenomena *are* at this time reference points for

explanations; people *do* have beliefs, intentions and so forth. If it is supposed that the present scope is on better ground than the earlier wide scope because the phenomena covered really are Intentional in virtue of being phenomena of goal-directed information processing systems, the

reply is that our notion of a goal-directed information processing system is part and parcel of the Intentionality in our conceptual scheme. A computer is no more *really* an information processor than a river *really* had desires.

What the purely extensional theory of behaviour would not say about

beliefs and intentions the extensional theory of the hydraulics of river flow does not say about the river's desire to reach the sea.

XI PERSONAL AND SUB-PERSONAL LEVELS OF

EXPLANATION: PAIN

The aim of Part I has been to describe the relationship between the language of the mind and the language of the physical sciences. Chapter1

proposed a stance of ontological neutrality with regard to expressions in the language of the mind, and since then these expressions have been seen to play a certain role in Intentional interpretations of certain physical systems, but how the ontological stance is supposed to mesh with the later developments has been only dimly suggested. The time has come to

consolidate the gains of Part I by illustrating them in application to a particular mental phenomenon: pain. Pain has been chosen because of its central role in a remarkably wide variety of philosophical and

psychological theories. Pains are the identity theorists' most plausible candidates for brain processes, but also in other theories the most

compelling examples of 'emergent' qualities or 'epiphenomena'. Pain, as we have seen, has a crucial function in stimulus-response behaviourism, but also figures centrally in the literature of the introspectionists and Phenomenologists.

The physiology of pain is relatively well understood. When a pain is

felt, neural impulses travel from the area in which the pain is felt along an anatomically distinct neural network for the transmission of pain stimuli.

In many instances there is a peripheral reflex arc that triggers withdrawal, but there are also other as yet unanalysed effects in the central areas of the brain. This, of course, is in harmony with the view of genetically

transmitted links developed in Chapter 3. It is appropriate for an organism to heed the most pressing demands of survival, and the imminence of

injury or death is as pressing as a demand can be, so it is altogether to be expected that a strongly entrenched pain network, essentially including appropriate responses of withdrawal, should be inherited. Moreover, as

personal experience reveals, the behavioural reactions to pain are more difficult to overrule than any other behavioural tendencies. Genuine pain behaviour is compulsive, involuntary, and only with great 'will power' or special training can man or beast keep from reactions to pain. Whether or not such inherited afferent-efferent networks are a *sufficient* condition for

the existence of the 'phenomenon of pain', it is safe to say they are a necessary condition. That is, it would be a very mysterious view that held that the bare phenomenon of pain could occur on the evolutionary scene

before there were organisms that reacted appropriately to stimuli that were harbingers of injury. Pain could not appear until organisms began avoiding it. The question before us now is whether pain is something (some *thing*) in addition to the physical operations of the pain-network.

An analysis of our ordinary way of speaking about pains shows that no

events or processes could be discovered in the brain that would exhibit the characteristics of the putative 'mental phenomena' of pain, because talk of pains is essentially non-mechanical, and the events and processes of the brain are essentially mechanical. When we ask a person why he pulled his hand away from the stove, and he replies that he did so because it hurt, or

he felt pain in his hand, this looks like the beginning of an answer to a question of behavioural control, the question being how people know

enough to remove their hands from things that can burn them. The natural

'mental process' answer is that the person has a 'sensation' which he

identifies as pain, and which he is somehow able to 'locate' in his

fingertips, and this 'prompts' him to remove his hand. An elaboration of this answer, however, runs into culs-de-sac at every turning.

The first unanswered question is how a person distinguishes a painful

sensation from one that is not painful. It is no answer to say that painful sensations are just those that hurt, for then the question becomes how a person distinguishes sensations that hurt from sensations that do not. If this question is seen as asking for a criterion for sensations that hurt, a criterion used by the person to distinguish these sensations, the question admits of no answer, for one does not distinguish the sensations that hurt or are painful by applying some criterion; one simply distinguishes them.

Their only distinguishing characteristic is painfulness, an unanalysable quality that can only be defined circularly. Moreover, a person's ability to distinguish this quality in sensations is ensured; one simply *can tell* when a sensation is painful (excluding cases where one's doubt is over whether the word 'pain' is too strong for the occasion). When trying to explain the discrimination of pains, appeal to the quality of painfulness is no advance over the question; it tells us nothing we did not already know. When one is asked how one tells an *x* from a *y* and answers that *x*'s have an indefinable

characteristic which one is simply able to recognize but not describe, all one is saying is: I can tell – that's all.

The *mechanical* question, how is it done? is blocked. It is blocked not because the reply is that one is in the dark about how one distinguishes painful sensations from others, but because the reply is that no mechanical answer would be appropriate in this context. Pains or painful sensations are

‘things’ discriminated *by people*, not, for example, by brains (although brains might discriminate other things related to pains), and the question is: how do *you* (the person) distinguish pains from other sensations? The question admits of no answer because the *person* does not *do* anything in order to distinguish pains; he just distinguishes them. Distinguishing pains is not a personal *actvity*, and hence no answer of the form, first I do A and then I do B, makes any sense at all. But if this is so, the appeal to a quality of these discriminated sensations is gratuitous. A quality, to do any work in a theory, must be identified, but this means it must either be described or ostended. Description presupposes analysis, and in this instance

analysis presupposes personal activity; where discrimination occurs

without personal activity, no description of a discriminated quality is possible. Then, if the quality is to be identified at all, it must be ostended, but ostension of the quality in this instance cannot be separated from

ostension of the discriminating. Where discriminating is an analysable

personal activity, like discriminating good apples from bad by checking for colour and crispness, we can distinguish the qualities from the

discriminating of them, but in the case of distinguishing sensations as painful, the act of discrimination itself is the only clue to the localization (in space and time) of the presumed quality. Insisting that, above and

beyond our ability to distinguish sensations as painful, there is the quality of painfulness, is thus insisting on an unintelligible extra something.

The first cul-de-sac, then, is that a *person’s* power to discriminate painful sensations is a brute fact subject to no further questions and

answers. The next question concerns the location of these pains, and meets the same fate. We do not locate our pains with the aid of any

independently describable qualities or ‘local signs’ provided us by the sensations; we just *can* locate them. Whatever the brain may be ‘doing’

when one locates a pain, the person does not do anything in the process of locating his pains, for there is no such process that a person could engage in. One could engage in the process of locating another person's pains, by

asking him questions, poking around until he screams and so forth, but not in the process of locating one's own pains.

The third question left unanswered has already been shown to have no

answer. What is there about painfulness that prompts us to avoid it,

withdraw our hand, attempt to eliminate it? The question is dead because there is nothing *about* painfulness at all; it is an unanalysable quality. We simply do abhor pain, but not in virtue of anything (but its painfulness). If, in our attempt to build an explanatory bridge between sensation and action here, we invoke the appreciation of an unanalysable quality of painfulness, we are forced to choose between two non-explanations. We can either take it as a contingent fact that painfulness is something we dislike, but a contingent fact that admits of no explanation since painfulness is

unanalysable; or we can take painfulness as necessarily abhorrent –

something which by definition we withdraw from or avoid – in which case there is no room for explanation since 'we avoid pains' is then analytic, and cannot take the 'because' of causal explanation after it.

When we have said that a person has a sensation of pain, locates it and is prompted to react in a certain way, we have said all there is to say within the scope of this vocabulary. We *can* demand further explanation of how a person happens to withdraw his hand from the hot stove, but we cannot

demand further explanations of terms of 'mental processes'. Since the

introduction of unanalysable mental qualities leads to a premature end to explanation, we may decide that such introduction is wrong, and look for alternative modes of explanation. If we do this we must abandon the

explanatory level of people and their sensations and activities and turn to the *sub-personal* level of brains and events in the nervous system. But when we abandon the personal level in a very real sense we abandon the

subject matter of pains as well. When we abandon mental process talk for physical process talk we cannot say that the mental process analysis of *pain* is wrong, for our alternative analysis cannot be an analysis of pain at all, but rather of something else – the motions of human bodies or the

organization of the nervous system. Indeed, the mental process analysis of pain is correct. Pains are feelings, felt by people, and they hurt. People can discriminate their pains and they do this not by applying any tests, or in virtue of any describable qualities in their sensations. Yet we do talk about the qualities of sensations and we act, react and make decisions in virtue of these qualities we find in our sensations.

Abandoning the personal level of explanation is just that: *abandoning* the pains and not bringing them along to identify with some physical

event. The only sort of explanation in which ‘pain’ belongs is non-

mechanistic; hence no identification of pains or painful sensations with brain processes makes sense, and the physical, mechanistic explanation

can proceed with no worries about the absence in the explanation of any talk about the discrimination of unanalysable qualities. What is the

physical explanation to be? Something like this. When a person or animal is said to experience a pain there is afferent input which produces efferent output resulting in certain characteristic modes of behaviour centring on avoidance or withdrawal, and genuine pain behaviour is distinguished

from feigned pain behaviour in virtue of the strength of the afferent-

efferent connections – their capacity to overrule or block out other brain processes which would produce other motions. That is, the *compulsion* of genuine pain behaviour is given a cerebral foundation. Now would this

account of pain behaviour suffice as an account of *real* pain behaviour, or is there something more that must be going on when a person is really in pain? It might be supposed that one could be suddenly and

overwhelmingly compelled to remove one's finger from a hot stove

without the additional 'phenomenon' of pain occurring. But although

simple withdrawal may be the basic or central response to such

stimulation, in man and higher animals it is not the only one. Could any sense be made of the supposition that a person might hit his thumb with a hammer and be suddenly and overwhelmingly compelled to drop the

hammer, suck the thumb, dance about, shriek, moan, cry, etc., and yet *still* not be experiencing pain? That is, one would not be acting in this case, as on a stage; one would be compelled. One would be physically incapable of responding to polite applause with a smiling bow. Positing some horrible (but otherwise indescribable) quality or phenomenon to accompany such a compelled performance is entirely gratuitous. [6]

In one respect the distinction between the personal and sub-personal

levels of explanation is not at all new. The philosophy of mind initiated by Ryle and Wittgenstein is in large measure an analysis of the concepts we use at the personal level, and the lesson to be learned from Ryle's attacks on 'para-mechanical hypotheses' and Wittgenstein's often startling

insistence that explanations come to an end rather earlier than we had

thought is that the personal and sub-personal levels must not be confused.

The lesson has occasionally been misconstrued, however, as the lesson that the personal level of explanation is the *only* level of explanation when the subject matter is human minds and actions. In an important but narrow sense this is true, for as we see in the case of pain, to abandon the personal level is to stop talking about pain. In another important sense it is false, and it is this that is often missed. The recognition that there are two levels of

explanation gives birth to the burden of relating them, and this is a task that is not outside the philosopher's province. It cannot be the case that there is *no* relation between pains and neural impulses or between beliefs and neural states, so setting the mechanical or physical questions off-limits to the philosopher will not keep the question of what these relations are from arising. The position that pains and beliefs are in one category or domain of inquiry while neural events and states are in another cannot be used to isolate the philosophical from the mechanical questions, for, as we have seen, different categories are no better than different Cartesian

substances unless they are construed as different ontological categories, which is to say: the *terms* are construed to be in different categories and only one category of terms is referential. The only way to foster the proper separation between the two levels of explanation, to prevent the

contamination of the physical story with unanalysable qualities or

'emergent phenomena', is to put the fusion barrier between them. Given

this interpretation it is in one sense true that there is no relation between pains and neural impulses, because there are no pains; 'pain' does not

refer. There is no way around this. If there is to be *any* relation between pains and neural impulses, they will have to be related by either identity or non-identity, and if we want to rule out both these relations we shall have to decide that one of the terms is non-referential. Taking this step does not answer all the philosophical questions, however, for once we have decided that 'pain'-talk is non-referential there remains the question of how each bit of the *talk* about pain is related to neural impulses or talk about neural impulses. This and parallel questions about other phenomena need detailed answers even after it is agreed that there are different sorts of explanation, different levels and categories. There is no one general answer to these questions, for there are many different sorts of talk in the language of the mind, and many different phenomena in the brain.

Part II

Consciousness

5

INTROSPECTIVE CERTAINTY

XII THE CERTAINTY OF CERTAIN UTTERANCES

The most central feature of mind, the ‘phenomenon’ that seems more than any other to be quintessentially ‘mental’ and non-physical, is

consciousness. In the chapters to follow, consciousness will be analysed from both the personal and sub-personal points of view, and the major

advantage to be gained from paying attention to possible sub-personal

accounts of consciousness will be that it will allow us to see that

consciousness is not *one* feature or phenomenon or aspect of mind, but several. Once the term ‘consciousness’ is seen to allude to an incompatible congeries of features, and these features are sorted out and described, many of the most stubborn perplexities in philosophy of mind dissolve.

The quest for a plausible and consistent analysis of consciousness

develops into the hunting down of that elusive quarry, the little man in the brain, who is driven first from his role as introspector only to reappear as perceiver, reasoner, intender and knower. Since Ryle’s *Concept of Mind*, we all scoff at the notion of this little man, but scoffing is not enough.

Expelling him from our thinking about mind requires, I hope to show,

more radical alterations in our views of mental phenomena than are

usually envisaged. It is one thing to exorcize the ghost in the machine, but he can reappear in more concrete form, as, for example, a stimulus-checking mechanism or – as we have seen – as a brain-writing reader, and in these guises he is equally subversive.

Our avenue to consciousness in ourselves is generally held to be the

faculty of introspection, and our avenues to consciousness in others are their introspective reports. Getting at the putative phenomenon of

consciousness requires that we first understand these modes of access, and the traditional problem with these is that they seem to be infallible in some strange way; we seem to have certainty about the contents of our

own thoughts.

The intuited commonplace that we cannot be mistaken about the content

of our own consciousness has been variously expressed and explained in

the philosophical literature. The picture, due ultimately to Descartes, of the introspector infallibly perusing the presentations of consciousness has been generally acknowledged as confused, but the alternatives proposed

have so far fallen short of giving a satisfactory account. The most

promising rivals to the Cartesian view all start from the observation that since any referring, factual report can be mistaken, our introspective

utterances must not be referring, factual reports. Thus Wittgenstein holds (or is often held to hold) that the invulnerability to error of *pain* reports is due to the fact that 'the verbal expression of pain replaces crying and does not describe it' – and hence is not a report at all, but akin to such other behavioural manifestations as writhing and crying. 1 Ryle adopts a similar position in *The Concept of Mind*, saying that reports of pain are 'avowals', not assertions. 2 Miss Anscombe's solution is to claim that pain reports and some other introspective reports are not cases where we have *knowledge* of what we say, but where we merely *can say* what we say: 'there is point in speaking of knowledge only where a contrast exists between "he knows"

and "he (merely) *thinks* he knows" '. 3 These views all have in common the move of making introspective reports the sort of things to which 'right'

and 'wrong' or 'true' and 'false' do not apply, but in a variety of ways they are implausible. When I tell the doctor the pain is in my big toe I am

certainly not just doing a sophisticated bit of whining, as Wittgenstein's view suggests, for I fully intend to *inform* the doctor. Ryle's view suffers from a parallel defect, and both views, however plausible they can be

made for reports of pain, become highly implausible when other

introspective utterances are considered. Anscombe's view is plausible

until one asks how she proposes to distinguish the fact that I *can say* all sorts of gibberish from the fact that I *can say* where my pain is. Her view depends on the sense of 'can say' which is the same as 'can tell', and 'can tell' reintroduces the notion of truth and the accompanying question of *how* we can tell. The answer to this question is that we just can, that's all.

In § 11 it was claimed that explanations in terms of pains and persons'

reports of pains *do* reach an abrupt halt at this point; there *is* nothing more to be said from this stance, but from another stance an explanation can be given of this primitive 'ability' we have.

These three views are on the right track in attempting to avoid the

Cartesian view of the infallible reporter, the impossibility of which can be seen by noting its analogical character. Since a reporter, a human being, can wrongly identify what he sees (what things are out there), merely

moving him 'inside' and making him an introspecting whatever-it-is is not going to ensure that he will infallibly report experiences (what things are in here). One cannot have reports without a reporter, so the notion of

infallible reports must just be wrong. Where the three views go off on a wrong track is in supposing that the solution can be given at the personal level of explanation. All three views deny, from the stance of ordinary mental language talk about pains, thoughts and so forth, that introspective utterances are – from this stance – what they so manifestly are: *reports* of pains, thoughts and so forth that can, like any reports, be true or false. The reporter of mental experiences is, as everyone knows, the *person* himself, and what he is doing is reporting, not moaning or avowing or engaging in a

sort of glossolalia to which questions of truth do not apply. We cannot answer the question of *how* these reports are infallible by denying that they are reports. If we are unsatisfied – as I think we must be – with an early end to explanation here, namely that introspective reports just are infallible, we must abandon the personal level and ask a different

question: how can introspective utterances be so related to certain internal conditions that they can be viewed as error-free indications of these

internal conditions? The relationship between this question and the earlier one is not at all obvious, but will become clear once an answer is sketched out.

At the sub-personal level, the key to the solution of the problem lies in the distinction between a functional or logical state of a system and a physical state. Putnam first pointed out the remarkable and fruitful

analogy 'between logical states of a Turing machine and mental states of a human being, on the one hand, and structural states of a Turing machine and physical states of a human being, on the other.'[4] Turing devised a general way of describing the organization of any computer or automaton in terms of an ordered collection of *logical states*, which are completely specified in a *machine table* by their relations to each other and to the input and output of the automaton, but whose physical realization in

'hardware' is left open. Any system for which a machine table can be

specified is a Turing machine, and a *particular* Turing machine (as characterized by a particular machine table) might be built in a variety of very different ways, e.g., directly out of electronic components, or

'simulated' in an existing computer, or with hydraulic valves and

plumbing, or by a large room full of people given certain tasks. Thus one identifies a Turing machine by the functional interrelation of its states, not by its physical constitution, and, similarly, a logical state is the state it is in virtue of its relations to other states and the input and output, not its

physical realization or characteristics. A particular machine T is in logical state A if, and only if, it performs what the machine table specifies for logical state A, regardless of the physical state it is in. Putnam explains: Now let us suppose that someone voices the following objection:

'In order to perform the computation [of the 3,000th digit of p]

just described, T must pass through states A, B, C, etc. But how

can T ascertain that it is in states A, B, C, etc.?'

It is clear that this is a silly objection. But what makes it silly?

For one thing, the 'logical description' (machine table) of the

machine describes the states only in terms of their *relations* to each other and to what appears on the tape. The 'physical

realization' of the machine is immaterial, so long as there *are*

distinct states A, B, C, etc., and they succeed each other as

specified in the machine table. Thus one can answer a question

such as 'How does T ascertain that X?' (or 'compute X', etc.) only

in the sense of describing the *sequence of states* through which T

must pass in ascertaining that X (computing X, etc.), the rules

obeyed, etc. But there is no 'sequence of states' through which T

must pass to be in a single state! Indeed, suppose there were –

suppose T could not *be* in state A without first *ascertaining* that it was in state A (by first passing through a sequence of other states).

Clearly a vicious regress would be involved. And one 'breaks' the

regress simply by noting that the machine, in ascertaining the

3,000th digit in p, *passes through* its states – but it need not in any significant sense 'ascertain' that it is passing through them.5

Suppose T 'ascertained' it was in state B; this could only mean that it behaved or operated *as if* it were in state B, and if T does this it *is* in state B. Possibly there has been a breakdown so that it should be in state A, but if it 'ascertains' that it is in state B (behaves as if it were in state B) it *is* in state B.

Now suppose the machine table contained the instruction: 'Print: "I am

in state A" when in state A.'6 When the machine prints 'I am in state A'

are we to say the machine *ascertained* it was in state A? The machine's

'verbal report', as Putnam says, 'issues directly from the state it "reports"; no "computation" or additional "evidence" is needed to arrive at the

"answer".' The report issues directly from the state it reports in that the machine is in state A only if it reports it is in state A. If any sense is to be made of the question, 'How does T *know* it is in state A?', the only answer is degenerate: 'by being in state A'. 'Even if some accident causes the printing mechanism to print: "I am in state A" when the machine is *not* in state A, there was not a "miscomputation" (only, so to speak, a "verbal slip").' Putnam compares this situation to the human report 'I am in pain', and contrasts these to the reports 'Vacuum tube 312 has failed' and 'I have a fever'. Human beings have some capacity for the monitoring of internal *physical* states such as fevers, and computers can have similar monitoring devices for their own physical states, but when either makes a report of such internal physical conditions, the question of how these are

ascertained makes perfect sense, and can be answered by giving a succession of states through which the system passes in order to ascertain its physical condition. But when the state reported is a logical or

functionally individuated state, the task of ascertaining, monitoring or examining drops out of the reporting process.

A Turing machine designed so that its output could be interpreted as

reports of its logical states would be, like human introspectors,

invulnerable to all but ‘verbal’ errors. It could not *misidentify* its logical states in its reports just because it does not have to identify its states at all.

If the analogy to human introspection is to be more than just suggestive, however, we must develop a more detailed picture of a machine which

makes ‘introspective’ reports.

XIII A PERCEIVING MACHINE

We want to describe a machine that would report its ‘mental experiences’

with the infallibility of human introspectors. Such a machine will require quite a sophisticated print-out capacity for making its reports, and, if the analogy is going to be convincing in detail, we must first consider how the human behaviour of speech might be controlled by neural mechanisms. It

would be naïve to suppose that introspective reports, or indeed any human utterances, are the immediate functions of any interesting internal logical states, on the model of Putnam’s machine print-out ‘I am in state A’. The production of speech is highly mediated by systems into which at present we have only meagre insights, but some general details of speech controls can be derived from an examination of the structure of language itself. The utterances of a natural language vary in certain rule-governed ways, and could only be produced by systems having certain sorts of organization.

Chomsky and others have initiated important research in this area, and one of the most important implications of their work is that the controls of linguistic behaviour must be hierarchically rather than serially arranged. [7]

There must be a control for the *whole* sentence or utterance that precedes and directs the production of each word or phoneme in turn. Applying the

loose notion of content ascription to these hierarchies, we can describe a hierarchy of *commands*. Last-rank efferent events hardly need be given content; their commands amount to ‘contract muscle’ or, slightly higher,

‘tongue forward’, and so forth. The commands organizing these would be

phonemic, ‘utter: “o” ’; at the next level up, events would control the organization of phonemic sequences. Here the command should not be in

the form of ordinary quotation (‘utter: “the cat is on the mat” ’), since, for example, ‘bear’ and ‘bare’ are phonemically equivalent and hence not

distinguishable at this level of control. In some cases of verbal behaviour the goal is merely the production of a phonemic sequence, as, for example, in beginning foreign language class drills, and the higher controls of these activities would be commands of the form ‘utter: …’ followed by a

phonemic sequence, and the only command of interest above that would be

‘mimic the teacher’ or something like that.

In slightly different cases such as taking an oath, reciting a poem or, in general, quoting someone or some document, higher commands would

have an *oratio recta* (direct quotation) content: ‘say: “I do solemnly swear

…” ’, and the overriding control might be given the content ‘recite what’s put before you’. The elaboration of these controls would, of course, differ in different people; the child may call out one word at a time while the adult reciter’s controls may govern the production of whole phrases or

sentences.

What is missing from these cases but is normally present in verbal

behaviour controls is a command with *oratio obliqua* (indirect quotation) content: ‘say that …’,‘ask him whether …’. Here, in contrast to the cases of recitation or quotation, what is to be done can be done in a number of different ways, what is to be said can be expressed variously. Not all such

controls would need to be given *oratio obliqua* contents, especially where what is to be performed is a speech act for which we have a name.

'Apologize' might be used as the content of an event at a level one step higher than the *oratio recta* commands 'say: "pardon" ', 'say: "excuse me" ', and 'say: "terribly sorry" '. It is tempting to go overboard at this stage and decide that the variations in event content at this level coincide with variations in our inner thoughts and on the basis of this proclaim the *identity* of thoughts with this sort of postulated brain process. For example, the event to which we gave the content 'apologize!' on the basis of behavioural effect might be adjusted on the basis of stimulus conditions or more central causes so that it was given, in one case, a content

coincident with the thought that one was genuinely sorry, and in another, a content coincident with the thought that protocol demanded an apology,

but we shall see that there are obstacles in the way of making such a

straightforward identification.

Projecting the structure of language in this way into the brain and

postulating hierarchical assemblies for controlling the production of

utterances is, of course, a rather empty trick. No actual mechanisms for doing this work and no discrete *anatomical* hierarchies are being proposed, but still this projection can provide a fruitful way of looking at things, of ordering the experimental and everyday observations we have about verbal behaviour. For example, we observe errors in verbal behaviour and can

now assign the malfunctions responsible for these to specific locales on the map of hierarchies, even though this does not in any real sense tell us where these malfunctions occur in the brain. Stuttering is the sort of

mistake we believe to be closest to the muscles, although we are prepared

to find much deeper causes for it; lisping and saying 'twee' for 'tree'

would be assigned to fairly stable bits of misprogramming at the

phonemic level. Above this would be the malfunctions responsible for

spoonerisms and other mistakes in phonemic sequence, then

mispronunciations and, still higher, strictly verbal and grammatical errors, including both ‘Freudian slips’ and more permanently entrenched misuses of words and malaprops. The arguments of Chomsky and others would

place responsibility for errors in syntactical ordering still higher, since the evidence suggests that the determination of syntactic structure is prior to word choice. Above this level there are only the solecisms that are not, strictly speaking, linguistic errors at all, like saying ‘Oops, dammit’

instead of ‘Please excuse me’. The particular difficulties that aphasics have in finding the word ‘on the tip of their tongues’ during post-stroke recovery (and in Penfield’s remarkable cases of electrically induced partial aphasia8) would be caused by malfunction at the level of *oratio recta* implementation of *oratio obliqua* commands.

Introducing the possibility of inhibition at the various levels, we can formulate a plausible sketch of what it is to talk to oneself, and to think.

Sometimes when we say we are thinking we mean in a very strong sense

that we are talking to ourselves: whatever is going on is expressed in full sentences, in definite words in a definite order, even in a particular ‘tone of voice’ with particular emphases, and the thinking of these thoughts

takes just about as long as saying the words aloud would take. But

sometimes our thoughts are not like this; sometimes they are swift,

somehow not quite formulated into particular words, and in no tone of

voice. In the former case one is tempted to agree with Ryle’s description of thinking: one is talking without moving one’s lips. One can talk, or one can whisper, or one can just move one’s lips; *or*, one can eliminate even the lip movements and whatever is left is this type of thinking. 9 Then one might describe the latter as further eliminating the formulation of

temporal sequences of particular words. Something of temporal succession remains but it is not the same as the easily clockable sequence of words in the former case. The physicalist supposition to go along with this

'introspective' account is that when one is talking to oneself the situation differs from when one is talking out loud in that the last-rank efferent impulses are inhibited, and that when the efferent activity is inhibited at

the level of *oratio recta* commands, the swifter form of thinking is going on.

Such a view is plausible, I believe, but it does not lead me to propose an identity of thoughts with these brain processes, even with these brain

processes assigned a certain content strikingly like the content we would normally assign to a thought. For immediately the objections would arise that no mechanism has been proposed to make me *aware* of these neural activities (and I certainly am *not* aware of these neural activities, while I *am* aware of my thoughts), and in any event the *content* of the activities is not at all a discriminable characteristic of them, such as I might be able to

'intuit', but merely an artificial determination made by some observing neurologist. These objections betray, I believe, a fundamental

misunderstanding of the problem, but they do hold against such a naïve

identity theory, which betrays the same misunderstanding. There is a lot more to be done before any sort of an answer should be attempted to the question of *what a thought is*, and if such a hypothesis about the organization of linguistic behaviour controls is likely to be a part of the answer, there is no reason yet to propose any identities. The hypothesis does give us a very general, Intentionally characterized model of the

organization of a 'speech centre', which is what we need for the

elaboration of our perceiving machine.

Suppose that the art of making neural net stimulus analysing

mechanisms has advanced to such a state that it was feasible, and desirable for some reason, to make a ‘perceiving machine’. Its sense organs could be television cameras (two for binocular overlap), and the output from

these cameras could be recoded in any regular way to fit the input

requirements of an immense neural net analyser which then fed *its* output into a ‘speech centre’ computer. The speech centre computer would be

programmed to transform the output of the analyser into printed English

‘reports’, like ‘I see a man approaching’.

It might be worth mentioning that there would be no need for television *screens* in this machine. Setting up the screens and then monitoring them with some device would simply postpone the activity of the analyser.

Since a television output, unlike the output of the eye, is in the form of a sequential stream rather than a simultaneous multi-channel barrage, it

would probably be advisable to ‘spread’ the sequence of impulses

reporting each complete scanning of the television camera image by time-

lags over a bank of inputs in the analyser, so that single scannings are fed in simultaneously, but this is a point of engineering, and not a logical requirement. Similarly, if one did arrange in this way for spreading the television output over an array of inputs, there could be only reasons of engineering (e.g., economy of wiring) for having the array reproduce the image in the camera. Since nothing will be *looking at* (or photographing) the arrays (no little man in our perceiving machine), there is no need for the pattern of inputs to produce any image or topological analogue of the sense organs’ image. Any stable spreading system could be used.

The analyser would eventually produce outputs to which one would

have to *assign* significance – by the arduous procedure of checking the multitudinous outputs against the vast variety of scenes set before the cameras and finding regularities between descriptions of the scene and

outputs. The trick would then be to programme the ‘speech centre’

computer to take over this job and produce English sentences describing the scenes presented as a transformation of the analyser’s output. Such a task is out of the question at present, but it is plausible to assume that the efficient way of programming the speech centre computer would be to

organize it along the hierarchical lines described above. Probably the only remotely feasible way to achieve this would be to build in certain

‘learning’ capacities in the speech centre computer and ‘teach’ it to

produce (true) English sentences. The ‘perceiving machine’ that resulted from all this miraculous expertise would, of course, be a pale copy of a human perceiver, since no provision would be made for it to use its

‘perceptions’ for any purpose other than as the basis for verbal reports, nor would the machine be given the capacities to lie about its view, to decide to talk about some other subject, to ask questions, etc. It would

simplemindedly reel off reports of what it saw – giving almost Skinnerian verbal responses to its visual stimuli. But it would share one crucial

feature with human perceivers: it could not be mistaken about its ‘mental’

states.

Once such a machine were operational, in what ways could its reports be fallible? First, it would be fairly easy to trick the machine. Presenting it with a moving dummy could result in the report ‘I see a man

approaching’, or, for example, the television outputs produced when a man was approaching could be recorded and then fed into the analyser at some time when there was no man approaching, producing something like an

hallucination in the machine, or one might say one had ‘hypnotized’ the machine. In these cases the analyser would issue in the same output as for veridical ‘perceptions’ of men approaching. Aside from such trickery

there might be malfunctions in the television system or the analyser. This could be guarded against by redundancy measures, but would still be

possible, we can suppose.

Such a malfunction or bit of trickery would result in an analyser output that was mistaken *relative to the outside scene*. This mistaken output would be expressed, in a false English sentence, by the speech centre

computer unless it too made a mistake. Feedback loops and redundancy in the speech centre computer would be designed to correct malfunctions

before the actual print-out, or if the malfunction took place in the last rank

– the actual printing – such typographical errors could be erased and

corrected by further feedback loops. If feedback loops failed to correct speech centre malfunctions there could be ‘verbal’ errors in the final

report, but if ‘verbal errors’ are discounted or corrected, whatever analyser output does enter the speech centre will be correctly expressed relative to the rules of language programmed into the computer. Disallowing misuse

of language and ‘slips of the tongue’ there is no room for mistakes to

occur in the expression of the analyser’s output. That output, whether right or wrong relative to the actual sensory scene, cannot help but be correctly expressed if feedback corrects all verbal errors. Our machine, like any machine, can malfunction at any point inside, but all the possible

malfunctions sort themselves into two kinds, depending on where they

occur. Uncorrected errors that occur prior to speech centre input are all errors in afferent analysis; any such error will ensure that the ultimate output of analysis (the input to the speech centre) is mistaken relative to what was being analysed: the outside world. Any uncorrected error

occurring after initial input into the speech centre will result in a verbal slip, a mistake in *expression*. Put another way, errors prior to speech centre input

make for errors in *what is to be expressed* by the speech centre, but if the speech centre functions properly, the only possibility of error is relative to the *outside* world. The key word is ‘expressed’. The perceiving machine as a whole can be said to make reports describing its external

‘visual’ environment, but it does not *report* or *describe* the output of the analyser, since that output is not a *replica* of what is outside, but a *report* of sorts itself. The speech centre part of our machine does not examine or

analyse its input in order to determine its qualities or even its similarities and dissimilarities with other inputs, but rather produces English

sentences as *expressions* of its input. The infallibility, barring verbal slips of the ‘reports’ of the analyser output, is due to the criterion of identity for such output states. What makes an output the output it is is what it goes on to produce in the speech centre, barring correctible speech centre errors, so an output is precisely what it is ‘taken to be’ by the speech centre,

regardless of its qualities and characteristics in any physical realization.

There are, then, two kinds of errors the machine can make. It can be

wrong in its analysis of what is actually before its ‘eyes’ (as in illusions, hallucinations and other misidentifications), and it can make only verbal errors in ‘uttering’ its reports of perception, but it cannot misidentify the output which comes from the analyser, which is the same logical state as the speech centre input. In other words, it cannot be mistaken about that which it *seems to see*. Suppose that instead of making its reports in the form ‘I see a man approaching’ it always wrote ‘I seem to see …’, or ‘it is just as if I were seeing …’. Reports in this form would disavow

responsibility for fraudulent input or mistakes in the analyser and hence would be infallible barring only correctable verbal errors.

This should not be taken to mean that the change in the form of words

changes what is going on; the switch from the ‘I see …’ idiom to the ‘I seem to see …’ idiom does not ensure that a particular thing is being done

(e.g., a report is being made about output rather than about the outside world), but that what is being done is to be interpreted in a certain way.

Whatever the form of words, whatever the sequence of printed symbols,

what is printed will be an expression of the analyser output; the form of words is just being used as an indicator that one is to discount

discrepancies between output and outside world. One could just as well

leave reports in the 'I see …' form and attach a small sign to the machine,

'Not Responsible for Fraudulent Input or Errors in Input Analysis'.

Carried over to the case of human utterance this point becomes: the

immunity to error has nothing to do with the execution of any *personal action*. An account of a man's intention, or of *what he thinks he is doing*, plays no role in explaining introspective certainty; with whatever *intention* an utterance is made (considered on the personal level), on the sub-personal level it will be an expression of the input into the human speech centre (which receives its input from more sources than just perceptual

analysers), and *as such* it is immune from error relative to the outside scene. In fact, of course, when we intend our utterances to be immune in this way, when we intend, that is, that others judge them in this light, we frame our expressions in the 'I seem to see …' idiom. In using this idiom a person is not intentionally expressing the input of his speech centre, for he has no notion of speech centre input at all, most likely; what accounts for the immunity to error is nothing the *person* does – no personal *action*, intentional or otherwise – but what is going on in his brain. 10

Using the notion of content ascription, and staying firmly on the sub-

personal level of explanation, we can say that a sentence uttered is not a *description* of a cerebral event, but rather the expression of the event's content, which, after all, may *be* itself a description – of the visual field, for example. As an expression, it is subject to verbal errors, but not to

misdescription or misidentification. 11 He who reads this sentence aloud is not uttering a description of the marks on the page, and although he may make a verbal slip, he cannot commit a factual error, since he is not

reporting or describing; for instance, lisping while reading aloud is not saying that there are *th's* on the page when in fact there are *s's*. The content of an event, or of the logical state of which a physical state is the

realization, is not a matter of intrinsic physical characteristics or qualities that could be reported or described, but of functional capacities, *including* the functional capacity to initiate (barring malfunction) just the utterance or class of utterances that would be said to express the content in some language. Thus the Intentional characterization of an event or state –

identifying it, that is, as the event or state having a certain content – fixes its identity in almost the same way as a machine-table description fixes the identity of a logical state. The difference is that an Intentional

characterization only alludes to or suggests what a machine-table

characterization determines completely: the further succession of states.

It is now even more tempting than before to identify thoughts and other mental events with certain other things, say Intentionally characterized brain processes given a certain functional location or logical states of the cerebral Turing machine, but still I will resist the temptation. The

argument that deflects me from this course is in some respects silly, but can be given enough force to satisfy *me* that there is no gain in proposing an identity and some danger of confusion. One could argue against the

identity that the mental experience or thought is what is *reported* when the

content of a certain cerebral state is *expressed*. For it is admitted that it is not the cerebral state that is reported and we *do say* that we report our thoughts and inner experiences, so a thought, being what-is-reported,

cannot very well be identical with what-is-expressed. (We do, of course, also talk of people expressing their thoughts, and this takes some if not all of the wind out of the sails of this argument. But we do not express a

neural event or state; we express its content, which is hardly the sort of *thing* one would go to the trouble of identifying with a thought.) Starting from the position that thoughts, being what-is-reported, cannot be

identified with anything in the sub-personal story, it would be poor

philosophy to argue further that there must really *be* something, the thought, that is reported when it is true that I am reporting my thoughts.

On that argument our perceiving machine would have to have thoughts, or at least thought-analogues, as well, and we have not instructed the

engineers to put thoughts in our perceiving machine. There is no entity in the perceiving machine, and by analogy, in the human brain, that would be well referred to by the expression ‘that which is infallibly reported by the final output expression’, and this is the very best of reasons for viewing this expression and its mate, ‘thought’, as non-referential. On the personal, mental language level we still have a variety of dead-end truths, such as the truth that people just *can* tell what they are thinking, and the truth that what they report are their thoughts. These are truths that deserve to be fused, and then the fact that there should be such truths can be explained at another level, where people, thoughts, experiences and introspective

reports are simply not part of the subject matter.

6

AWARENESS AND

CONSCIOUSNESS

XIV THE ORDINARY WORDS

The account of introspective certainty given in the last chapter is the first step in a theory of consciousness or awareness. The infallible reporter in the

mind has evaporated, to be replaced at a different level of explanation by the notion of a speech-producing system which is invulnerable to

reportorial errors just because it does not ascertain and does not report.

This is just a first step, however, for there are more aspects of

consciousness than just perceptual consciousness, and more things we do with speech than just make sincere reports of our experience.

This chapter will be an examination of our concepts of consciousness

and awareness with a view not merely to cataloguing confusions and

differences in our ordinary terms but also to proposing several artificial reforms in these terms. It is fairly common practice to use ‘consciousness’

and ‘awareness’ as if they were clearly synonymous terms, or at least

terms with unproblematic meaning, but I shall argue that these concepts, as revealed in the tangled skein of accepted and dubious usage, are an

unhappy conglomerate of a number of separable concepts and that the only way to bring some order and manageability to the task of formulating a

theory of consciousness and awareness is to coin some artificial terms to reflect the various functions of the ordinary terms.

The first thing to notice about the two words is that both of them have Intentional and non-Intentional uses. On the Intentional side, we speak of being conscious *of* this or that, aware *of* this or that, aware *that* such and such is the case, and – less naturally – conscious *that* such and such is the case. On the non-Intentional side, we speak of being just plain conscious or unconscious, and of being a conscious form of life, and, in rather

artificial speech, of someone’s simply being aware, in the sense of being

‘on the *qui vive*’ or sensitive to the current situation. We also speak of conscious and unconscious motives or desires, but these can be

assimilated under the Intentional idioms, as motives and desires we are conscious *of*.

Since ‘conscious that’ is at least unusual if not outright one of those things we ‘do not say’, and since ‘conscious of ’ and ‘aware of ’ are as close to being synonymous – to my ear – as any terms we are apt to find in ordinary language, a step in the direction of clarity and order can be taken

by abandoning ‘conscious that’ and rendering ‘conscious of ’ always as

‘aware of ’, thus forming all the Intentional idioms with ‘aware’. Then if it can be agreed that the non-Intentional use of ‘aware’ (as in ‘the younger generation is so aware!’) is just a fancy way of speaking of alertness

(‘heightened awareness’), it can be subsumed under the non-Intentional

sense of ‘conscious’, where it means, roughly, ‘conscious to a high degree’

– whatever that means. The move, then, is to group all and only the

Intentional senses of our two words under ‘aware’, and all and only the non-Intentional senses under ‘conscious’, and the excuse for the move is that no real violence is done to the variety and fragile subtlety of common talk, and a great gain in order is achieved.

Now there is no special reason to suppose that when we speak of

someone being aware of something (or, to use the abandoned idiom,

conscious of something) we are speaking at all to the subject of that

person’s being just plain conscious – in a non-Intentional sense. There may be something like an implication involved, such that in order to be aware of something one must be conscious, but this does not require that

awareness (always *of* something) is the same phenomenon as

consciousness. To keep the issues as separate as possible, I shall treat the two terms one at a time, examining awareness first, and returning to

consciousness later in the chapter.

There are two quite different sorts of features of situations that govern our talk of awareness. One of these is our dependence on awareness of

things for manoeuvring in the environment. It is this feature we allude to when we say ‘he *must* have been aware of the tree, for he neatly swerved around it’. A little girl darts out into the street; if I do *not* apply the brakes, there may be at the inquest some question about whether I was aware of

her, but if I *do* brake, my awareness of her will not be at issue, but only, say, *when* I became aware of her, whether I was speeding, and so forth. If we are tempted to say that dumb animals are aware of things, it is in virtue of their reactions to the environment: the cat springs but the bird flies away, so we suppose that at the last instant the bird became aware of the cat; the bee diverts from its bee-line to avoid collision with a tree, so the bee was aware of the tree. This raises a difficulty, however, for ‘aware’ is Intentional. Was the bird aware of the cat *as* a cat, or just as a hulking presence or even just as danger? Was the bee aware of the tree *as* a tree, or just as an obstacle? The latter alternative suggests oddly that the bee could

be aware of the tree in a rather sophisticated, *abstract* way as an obstacle, the way a man might view a tree merely as something to be got around.

One might say it is at least certain that if bees are aware of things at all they are not aware of things in a way at all like the way people are aware of things. One is tempted to add: we cannot *know* what bees are aware of; if only they could tell us!

We are tempted to say this by the other feature governing our talk about awareness: our ability to make introspective reports. The reason we feel safer in ascribing awareness of things *as* certain things to people is that they tell us. We do not know what it is like to be a bee or a bird, but we know what it is like to be blind or myopic or to have tunnel vision,

because people suffering from these conditions can describe their

experiences. The human capacity for making introspective reports is seen as a mode of access to the content of awareness, and in virtue of the

invulnerability to error examined in the last chapter, its deliverances are seen as reliable – indeed conclusive – evidence of the content of

awareness. We ask the driver if he was aware of the little girl, and his reply ‘No, all I was aware of was a swift blur of motion’ settles the content of his awareness provided we do not doubt his sincerity.

The difficulty is that these two features – behaviour control and

introspection – do not always mesh as we would like. A man may have

driven for a hundred miles, and when we ask him what he was aware of

along the way, he may reply ‘Nothing, since the route was familiar and I was engrossed in conversation with my passenger.’ If we allow his account to stand, then we must admit that awareness and behavioural control are separable, for the fact that he successfully steered the car around dozens of curves will have to be viewed as no evidence that he was aware of them.

Alternatively we may bully him into admitting that he was aware of the

curves (‘I guess I must have been aware of them in an abstract or vague sort of way’), but his acquiescence in this amounts to the abandonment of his status as authoritative introspector. The man’s ability to speak and more specifically to report his experiences counts for nothing over the muteness of the bee if his accounts are overruled on behavioural grounds.

There *is* an activity which is giving error-free introspective accounts of awareness, but it can be subverted by a misplaced allegiance to the other feature of awareness that interests us: behavioural control. Consider the man who reasons thus: I must have been aware that the glass had reached

my lips, or I wouldn’t have tipped it. This man is not introspecting. He is speculating, framing a hypothesis on no more evidence than any other

observer might have. He should have said: I was not aware of the glass at all; I was listening attentively to the conversation, and so cannot provide any *privileged* information on the perceptual cues that must have initiated my drinking. There are times when what we are aware of (in the sense of what we can introspectively report) is also what is relevant to our

behaviour; there are times when what we are aware of has nothing to do

with our current behaviour; and there are even times when *becoming* aware of what is directing our behaviour encumbers that behaviour. It has been shown that table-tennis players rely on the sound of the ball striking the table even more than on the sight of the ball. Suppose a table-tennis player said: I had no idea I was aware of the sound – except as a meaningless din

– but now I see I must have been, all along. Having said this he would

probably *start* being aware of the sound as more than a meaningless din and then his game would suffer, just as the typist or pianist who pays

attention to his finger motions becomes all thumbs.

Awareness sometimes seems to be a necessary condition for the

successful direction of behaviour, and yet in another sense awareness is clearly detachable from behaviour control, with some constraints (there are some limits on just how much of what one is doing one can be unaware of). When we say of the driver that he must have been aware of the curves under some description we are relying on the former sense of ‘aware’, and when the driver replies that he was conversing or daydreaming and

unaware of the curves he is relying on the latter sense, and the crucial point is that both we and the driver can be right at the same time. These two notions of awareness are entirely distinct in spite of their customary merger; what one can report directly, infallibly, and without speculation or inference is one thing, and what serves, or is relied upon, to direct

behavioural responses is another. This can best be brought out by coining artificial terms to mark the difference and observing the way these terms

behave when applied to different ordinary situations.

The first step in the baptism of the new terms is to recast all ‘aware *of*’

contexts into the ‘aware *that*’ context of propositional attitude, a move that harks back to the practice – and reasoning – of Chapter 2. For example, ‘I am aware of an apple on the table’ becomes ‘I am aware that there is an apple on the table’. Subtle differences in the normal sense of

these two expressions need not deter us; what is true in each case is that one is in receipt of a perceptual report (true or false) to the effect that there is an apple on the table. This use of ‘aware that’ is emphatically not intended to be the ordinary use, from which it differs in several respects.

First, the ordinary use is at least often truth-relative, like ‘know’: ‘are you aware that *p*?’ implies that *p* is true. Second, ordinary ‘aware that’ is usually used in a way that need have nothing to do with current

experiences or present state of consciousness at all; ‘are you aware that he is a judge?’ is not normally taken to be asking about what is running

through a man’s head at the moment, but rather what he knows – no matter what his attention is on at the moment. Setting these ordinary connotations aside and using ‘aware that’ simply to get everything into the guise of propositional attitude, the two senses of ‘aware’ can be defined:

(1) A is aware1 that *p* at time *t* if and only if *p* is the content of the input state of A’s ‘speech centre’1 at time t

(2) A is aware2 that *p* at time *t* if and only if *p* is the content of an internal event in A at time *t* that is effective in directing current behaviour.2

These definitions bridge the gap between the personal and sub-personal

levels of explanation. The ordinary personal-level term ‘aware’ is being replaced by two terms that still take persons (or whole systems) as

subjects, but have sub-personal criteria. I am proposing to explain the ordinary word ‘aware’ by abandoning it altogether and talking about two

very different words, ‘aware1’ and ‘aware2’. To some this may seem like an admission of madness, but there is no alternative method of analysis in this case. It is not the case that there are two clearly different *ordinary* senses of the word ‘aware’, the way there are two senses of ‘feel’ (‘I feel dizzy’ and ‘He is feeling his way around in the dark’). If there were, there would be no problem. Rather, what I have called the two senses of ‘aware’

are mingled and confused in our ordinary use of the term, so that, for

instance, when we say the dog is aware of the bone we think we are saying just the same thing about the dog as we say about the man when we say

that he is aware of the bone. It is this confusion that leads to most of the problems about awareness or consciousness. When we suppose, on the

basis of our casual observation of behaviour, that dumb animals are aware of various things, and when we wed this supposition to our personal

experience of awareness, we are left with the problem that *if* dumb animals are aware of things, have conscious experience, we can never

know what it is like, since they cannot tell us. In supposing that the

awareness we posit on the basis of clever behaviour is at all *like* human awareness of the sort we make introspective reports about, we only follow the actual, ordinary paths laid down by ordinary usage, but in following these paths we are led to error and confusion. Recognizing this confusion we may decide, as in fact some philosophers have, that dumb animals are not aware of anything, but this goes just as much against the grain of

ordinary usage. To say of a man that he is not aware of anything is to

suggest that ‘his mind is a blank’ and moreover that he is stumbling

around most ineffectually, and this is not what we wish to say about

animals. We do talk of an animal ‘being aware of every move we make’,

and *in some sense* we are right, but not in any recognizably separate *ordinary* sense of ‘aware’.

Given the two new definitions of awareness, it is at once clear what one could mean by saying that animals are not aware of things in at all the same way people are: animals are only aware2 of things, which is saying very little, since nothing in our definition would prevent certain cybernetic machines from also being aware2 of things. People are aware2 of things, but they are *also* aware1 of things, a possibility ruled out in the case of dumb animals. The temptation lapses to say we cannot know how animals

are aware of things, and if only they could tell us. If animals could tell us, they would be aware1 of things, which is entirely different. The concepts of awareness1 and awareness2 are distinct, and it is only when the halo of intuitions around one of these merges with the halo of intuitions around the other, as it does in our ordinary word, that confusion results. One can say, using these new definitions, that insects and birds and such are simply not aware1 at all, and then the question cannot be asked: But if the bee was not aware of the tree, how did it know enough to fly around it? The bee was aware2 of the tree. The driver who ‘paid no attention’ to the route was not aware1 of the curves (as curves or as anything else), but he was aware2

of them as curves. Consider the following exchange: ‘Fred, you haven’t

heard a word I’ve said!’ – ‘Quite right, Ethel. All I heard was an incessant

babble of noise.’ What was Fred aware of? There are several possibilities.

Had he been truly paying attention to her, Fred would have been aware1 of what she said, and also, of course, aware2 – as his responses would

indicate. Or he might have been aware1 of her talk only as a babble of

sound (like listening to a babbling brook), or even entirely unaware1 of her noise-making (as of the clock ticking). In the latter two cases he may have still been aware2 of enough to grunt ‘Oh really?’ occasionally at

appropriate moments. The fact that the ‘Oh really?’ was inserted at

syntactically appropriate moments would suffice to show that Fred was

aware2 of the talk as *more* than a meaningless din of babble, even if he was not aware1 of the sounds at all. The intermediate case of being aware1 of the sound as babble would in fact be highly unusual – a case of sitting back and adopting an aesthetic attitude to the musical tumble of syllables without being aware1 of their meanings. No doubt this is not what Fred

meant to suggest; most of the time he was probably entirely unaware1 of Ethel, but aware2 of the sentences to which he responded occasionally

'unconsciously'.

I do not want to suggest that the concept of awareness2 is tailor-made

for the use of investigators in animal behaviour and the brain. They may devise much more useful and rigorous concepts. The chief value of

'aware2' is in putting off those who would insist that awareness is a

prerequisite of regular, appropriate behaviour, and hence confuse the two functions of the ordinary word. The term is at least harmless, which is a step in the right direction. The more important concept, awareness1, is restricted to creatures that can express, or in other words, speaking

creatures. Non-speakers can no more be aware1 of things than be guilty of mispronunciation. Of course any machine that, like our perceiving

machine, had a speech centre attached would be aware1 of the content of the input to this speech centre, and this may seem to be an intolerable situation, but only if one clings to the folklore that has accrued to the ordinary word 'aware'. There is certainly nothing wrong with a machine

being aware1 of certain things if *all* this means is that it can express these things correctly. But if that is all the word means, there is still a great deal to explain or explain away, for I wish to show that there is no important

residue in the ordinary concept of awareness that is not subsumed under either awareness1 or awareness2. There is no room, I wish to show, for a concept of awareness3, which would apply only to people and rule out all imaginable machines.

XV AWARENESS AND CONTROL

The remarkable running together of the two notions of awareness in our

ordinary concept is perhaps to be explained by the fact that there is a high degree of coincidence of the two in human affairs. To a great extent we are aware1 of those of our activities we are most concerned to control well.

We *pay attention* in order to *do better*, but is there anything in the concept of awareness1 to suggest why this should work?

In § XIII we drew an imaginary line dividing the perceiving machine

into its two functional parts, the afferent analyser and the speech centre.

Let us call this line in any analogous system the *awareness line*. In the case of a manufactured machine it is plausible to suppose that we could draw a neat physical line separating the two stages, but it is much less likely that we could draw the analogous line in a person's brain. It would probably be gerrymandered out of all intelligibility. That would not,

however, diminish the importance of the *theoretical* line, and there are even a few guidelines that could be followed in an attempt to plot it. It has been suggested that feedback loops serving to correct malfunctions at each level in the production of utterances would be present in the speech centre part of the system. Such loops could not extend back into the analysis part of the system – for what standard would there be for them to test against?

So if such feedback loops could be anatomically distinguished (and this would be no mean trick, for we cannot expect these 'loops' to be anything but extremely complex and possibly ephemeral structures), once one had

reached the end of the feedback hierarchy (and this might not look

spatially like a hierarchy), one would have reached the ‘edge’ of the speech centre, the awareness line. Like the Equator, the awareness line is not itself a physical feature but rather a conceptual line projected on to a physical system. Once one has the concept of a great circle equidistant from the Poles one can determine the location of the Equator, which is not arbitrary or conventional, and the same holds for the awareness line in principle. Given the interlocking definitions of what it is to cross the awareness line and what it is to be aware1 of something, the sentence ‘one is aware1 that *p* when *p* is the content of an event that crosses the awareness line’ is analytically true, but not trivial. Before the Equator was

mapped ‘one is in the Southern Hemisphere once one has crossed to the south side of the Equator’ was uninformative in just the same way. The

definitional circle is broken once one provides an independent

characterization of the line, but one need not do this before making use of the concept.

In a person, if a signal (an event given content) does not cross the

awareness line, the person cannot express its content; he is not aware1 of its content. Such a signal, however, could contribute to some very useful circuit in the nervous system. A simple reflex like blinking or pulling one’s finger away from something hot is apparently controlled by

relatively short neural arcs that one would not expect to involve speech centre activity. Our experience shows us that by the time we become

aware1 that our finger has touched something hot the reflex has already occurred. One *can* become aware1 of one’s blinking, but one almost never is, and awareness1 is certainly not a necessary step between stimulus and response in this case.

It is not only simple reflexes that can apparently be controlled without the intervention of awareness1. An accomplished pianist can play difficult music beautifully ‘with his mind on something else’, and need not be

aware1 of the notes on the page, the sounds of his playing or the motions of his hands and fingers. He must, of course, be aware2 of these. There is nothing particularly remarkable about this, for could one not build a

machine that read music and then played it (a sophisticated player-piano)?

There is no temptation to suppose that *it* would be anything more than aware2 of what was going on. Experience suggests that although we can

only be aware1 of one thing at a time, the brain can control a number of complex activities at the same time. As we say, we do many things without thinking about them, but surely we do not do these things without the

brain's controlling them. It would be rare for a man to drive long distances without occasionally being aware1 of his driving or the landmarks, and

similarly the pianist would not long remain unaware1 of the notes, the

sounds or his finger motions. In particular, if he made a mistake, some sort of 'negative feedback' would no doubt shift him to awareness1 of

what he was doing.

This suggests that awareness1 does have some efficacy in behaviour control. What is the point of paying attention if not to control one's

behaviour better, and does not paying attention involve awareness1 of what one is doing? But at the same time it is clear that simply bringing a signal across the awareness line, 'into speaking position' so to speak, could have in itself no beneficial effect on behaviour control. There can be no logical relation between being aware1 of something and improving one's control

of related behaviour, but there could be a contingent and coincidental

relation.

We bring activities into awareness1 to correct them or improve them.

The pianist who keeps fumbling a trill starts paying attention to the

particular motions of his fingers when trilling. When learning to drive one is very much aware1 of raising the clutch, shifting gears, looking in the mirror and so forth, although these activities eventually become

'automatic'. It is also clear that we are inevitably aware1 of the sights, noises and other sensations (to use the word in its most ordinary sense) that are particularly bright, sudden, acute, bizarre, unexpected or otherwise outstanding. It would be appropriate for us to be aware1 of these if our awareness1 contributed to better coping with the environment, since the outstanding sensations are usually the ones that make the most difference to the person's well-being. Could it be that becoming aware1 of these

crucial events is a contingent but perfectly natural by-product of some shift in controls that occurs on the far (inner) side of the awareness line?

There seem to be two levels from which we direct our behaviour. At the

'high' level (apparently in the cortex) we correlate information from a variety of sources, the behaviour controlled is versatile and changeable –

and not particularly coordinated. Once under control, the behaviour is

often made into a routine and the control is packed off into a more

automatic and specialized system. (Apparently the site of these controls is the cerebellum.) If 'paying attention' is a matter of dealing with the

relevant parts of the environment at the high level, it might also *happen* to be a matter of bringing certain high-level signals across the awareness line, just because that is the way the brain is wired. We can even suppose that such connections would have survival value in that they essentially contributed to the human activity of (verbal) teaching and learning.

By viewing the relation between concentration or paying attention and awareness1 as not only contingent but exceptioned, we can account for

phenomena not otherwise describable without confusion. We can say, for

example, that dumb animals can pay attention (e.g., the first step in

training a dog is to teach him to pay attention to his master) without

sliding from an acknowledgment of this obvious fact into the supposition that in that case dumb animals are aware just the way people are. Or

consider a trained seal balancing on a ball while balancing another on its nose. A seal has a cerebellum and perhaps the seal, like a man, puts part of its trick on automatic pilot in the cerebellum, while *concentrating* on the other. We can investigate such possibilities without imagining that the seal, in concentrating, is rehearsing thoughts in seal language in his head.

Lacking a speech centre the seal cannot be aware1 of anything, cannot

introspect. This does not mean the seal is doomed to an unhappy life of uncomprehending darkness, always wondering what was going on both

inside and out. Only a being that can be aware1 of something can be sadly unaware1 of anything. People struck blind are depressed by the loss;

blindness does not bother stones.

Another case on which light is shed is that of pain. A sore foot may be so sore that I cannot ‘get my mind off it’, and am almost continuously

aware1 of the pain. Or I may be only intermittently aware1 of the pain. In the latter case I may or may not be able to concentrate usefully on other things during the intervals between awareness1 of the pain. Philosophers are fond of asking whether there are unfelt pains. The answer that suggests itself is that being in pain, for a person, is a dual situation, involving both awareness1 and awareness2 of the pain. I might not at every moment ‘be

conscious of ’ the pain (be aware1 of the pain), and yet the continuing neural excitation might disrupt the high-level operation of the brain and hence indirectly *bother* me, regardless of what I happened at any moment to be aware1 of. Not being always aware1 of the pain, I would not always have unimpeachable authority as to whether the pain was bothering me

(saying that it was the pain that was preventing me from working would be a hypothesis for which my only evidence would be the intermittent

awareness1 of the pain). When I ceased to be even aware2 of the pain we would say the pain had ceased altogether.

I think we should resist the temptation to choose one of the two new senses of 'aware' as *the* sense of the term; neither one can claim a clear majority of supporting intuitions drawn from our ordinary language. If

what is held to be essential to awareness is its relation to behavioural control ('he *must* have been aware of the tree, for he swerved') then it must only be an exceptioned coincidence that one can infallibly introspect what one is aware of; if what is held to be essential is this infallibility of expression ('only *I* can tell for sure what I am aware of ') then it is only usually true that we are aware of the information of foremost importance to the control of current behaviour. It is easy to confuse awareness1 with awareness2 and extrapolate from the fact that one must be aware of

something in order to say it (express it, report it) to the untruth that one must be aware of something in the same sense in order to do anything with it. For example, Sayre says that 'we would not say under any ordinary

circumstances that we recognized an apple, or some other object, but were aware of no such object'. 3 Of course I could not *say* that I recognized an apple and yet was not aware of an apple, since in order to say anything I must be aware1 of it, but this should not lead to the conclusion that *I* cannot say *he* recognizes an apple without implying that he is aware of the apple in the same sense. This illicit move is obscured by the use of 'we' in the quotation above. What of sorting machines or animals that cannot say anything? They can fulfil all the functions of recognition short of saying they recognize; does this bar them from recognizing? It is not that one must be aware1 in order to recognize, but that one must be aware1 in order to say that one recognizes. 4 To the extent that our ordinary concept of awareness leads to such confusions it is a poor concept in spite of its ordinariness.

XVI CONSCIOUSNESS

In § XIV, ‘conscious’ was found to have both Intentional and non-

Intentional uses, and the Intentional uses were subsumed under the

Intentional uses of ‘aware’. We distinguished the question of what it is to be conscious *of* something from the question of what it is to be conscious, and although this initial distinction allowed an account of the Intentional sense of ‘conscious’, we are not through making distinctions, for even in its non-Intentional uses ‘conscious’ is ambiguous. As Scriven for one

points out, ‘conscious’ can mean something like (1) ‘awake’ or ‘aware’ (in the non-Intentional sense of ‘alert’), or (2) having the capacity to be awake or aware. 5 That is, in sense (2) ‘conscious’ is used to distinguish beings capable of consciousness in sense (1) from inanimate objects. It is not paradoxical to say, then, that only conscious beings can be unconscious.

This is not an unusual ambiguity; only rational creatures can be irrational, only seeing beings can truly be called blind. Since the question of what it is to be conscious in sense (2) is entirely dependent on what it is to be conscious in sense (1), we can skirt this ambiguity by restricting our use of the word to sense (1). Once we know about it, we will be able to say with no further ado what it is to be conscious in sense (2).

Consciousness seems itself to be a capacity that comes and goes in

beings that have the capacity to be conscious, but what capacity is it? At first glance it seems to be the capacity to be conscious *of*, or aware *of* things, but is this awareness1 or awareness2? This question is implicit in the alternatives, ‘awake’ or ‘aware’. When a dumb animal is awake, it can be aware2 of things; is it then conscious? We have few qualms about

saying a dumb animal in a coma is *unconscious*, but what is its state when it is not asleep or in a coma? Is our reluctance to call animals conscious just a matter of confusing aware1 with aware2? Where do we draw the line between conscious and unconscious in any case? Is a person unconscious

when he is asleep or only when he has ‘passed out’ or is in a coma?

So long as we fix our eyes doggedly on the ordinary word ‘conscious’ and consider all that has been said in the past by philosophers and psychologists about consciousness, it is easy to convince ourselves that ‘conscious’ and ‘unconscious’ sunder the universe in a very fundamental way, and from this it is easy to arrive at the conviction that being conscious must be an all-or-nothing matter. Capacities can be partial, however, and admit of degrees (consider the common but ill-defined term ‘semi-conscious’). If being conscious is held to be having the present capacity to be aware2 of things, then when one is in a coma, one is unconscious to a very great degree, and when one is asleep one is unconscious to a lesser degree, since reflexes still work and one can wake up, which is itself a behavioural response to incoming information. If this is what we should mean by ‘conscious’, then animals without the power of speech can be conscious and people can even be aware of things while (relatively) unconscious. Consider a person who said, in describing a dream, ‘Suddenly I was aware of a man with a knife, and that scared me so much I woke up.’ The dreamer could be aware2 of the man in the dream; if he talked in his sleep he might also be aware1 of the man in his dream – and all this while asleep or only partially conscious.

If, on the other hand, consciousness is seen more restrictedly as the capacity to be aware1, of things, dumb animals can never be conscious, and dreamers and people who talk while in hypnotic trances are conscious.

I do not believe that an analysis of our ordinary language here will reveal that we mean one of these rather than the other. Ordinary language does not determine which of these alternatives is right, because it mixes them together. A decision on ‘conscious’, therefore, cannot be a decision about what the word actually means, but only what it *ought* to mean, and this can only be relative to our purposes. Most of the interesting theoretical

questions about consciousness seem to tie it to the notion of awareness1.

This is particularly clear in the case of the terms ‘subconscious’ and the Freudian ‘Unconscious’.

The control of reflexes in man is subconscious, as are the stages of

perceptual analysis, and in fact all information *processing*. We are not aware1 of the processes at all (as one might, with suitable incisions and mirrors, be aware1 of one’s digestive processes). What is subconscious, clearly, is everything that happens in the brain except what crosses the awareness line. As Lashley says, ‘No activity of mind is ever conscious.’[6]

He gives an example: responding to the request to think a thought in

dactylic hexameter. We are conscious of the thought we produce, but not of its production, or of *how* it was produced. The consciousness Lashley is

concerned with is clearly the capacity for awareness1; we *cannot say* how we produce this thought, and have no introspective access to the activity or process. Similarly, the Freudian Unconscious, if it is anything at all, is a region inaccessible to awareness1, and has nothing directly to do with

comas and sleep.

This proposed treatment of ‘conscious’ – and its brethren,

‘unconscious’, ‘subconscious’ and ‘Unconscious’ – completes the

fragmentation of the ordinary words of the chapter title. The

philosophically trained reader is apt to feel uncomfortable about this

analysis for two reasons. First, because of his recognition of the privileged position of ordinary language, he is apt to conclude that I have not so much *analysed* the ordinary words as *tampered* with them, an error that puts me in danger of speaking nonsense or at least irrelevancies. Second, he has been taught, by Ryle and others, to avoid the 'bogy of mechanism', and my analysis is accompanied by the spinning out of some unabashedly

mechanical and quasi-mechanical speculation. The two philosophers'

rules, 'Tamper not with ordinary words' and 'Avoid mechanism', are good rules, but their sound application is not universal. I would like to defend my particular transgressions of these rules.

In Chapter 1, I proposed as a *modus operandi* that sentences using the difficult 'mental' words be treated as significant, true and false, but not automatically subject to the sort of semantic analysis that generates

ontology. The question of whether 'thought', for example, refers to

anything was to be left for the time being up in the air. In Chapter 4 this hands-off policy was invoked when the questions were considered: *how* do we tell pains from non-pains, and *how* do we locate pains? These questions, framed in the language of people and pains, not bodies and

neural events, have no answer, or only the brusque answer: we just do,

that's all. Different questions may be asked and answered, but these are not strictly speaking *about* pains. If one is not confused by the brusque answer to the original questions, then the concept of pain is not confusing, is not to be criticized or discarded. Similarly, in Chapter 5 the temptation to identify thoughts with certain brain processes or their contents was resisted since we *do say* that we report our thoughts, and do not run into any *ordinary* confusions in saying this. No reason was seen for abandoning the idiom 'report one's thoughts' in favour of something clearer. In these instances the requisite correlations between the mental-language sentences

and physical-language sentences could apparently be made without first doctoring up the mental-language sentences.

With ‘conscious’ and ‘aware’, however, the situation is different.

Ordinary usage of these words is not remotely consistent. We *do say* that both people and animals are aware of things, that they are conscious or unconscious, that one can say what one is aware of or conscious of, and that one must be aware of something in order to recognize it. We do say these things, but we say them, even *ordinarily* (when not engaged in philosophical discussion) with misgivings. Our intuitions conflict when we are confronted with the crucial test cases. It is not merely that

philosophers can generate confusions by misusing these words, but that

the words in their most time-honoured uses are confused. Order can be

brought out of this chaos, but only by abandoning the conviction that

ordinary usage here is conceptually sound, that it meets the standards that are met by most words as a matter of course. And as soon as we abandon

this conviction we can no longer rely on the totality of usage to tell us what lies behind our notions of awareness and consciousness, because

nothing consistent lies behind the totality of usage. When we look for

distinctions that will serve to mark off consistent, separable senses of these words all we can find are distinctions of function, and pinning down these distinctions of function involves drawing a plausible mechanistic picture. There is an alternative – a ‘para-mechanical’ picture, to use Ryle’s term, replete with non-physical, non-causal ‘bits of not-clockwork’.

Avoiding this alternative, one is still forced to do psychology rather than

‘pure’ philosophy.

The psychological hypotheses, however, do not miss the mark when

applied to the philosophical questions. Can a machine be conscious? This question cannot be answered until we arrive at a conclusion about what it is to be conscious, and ordinary language does not tell us. Consider the

philosophical puzzle of whether or not animals can be conscious and the related puzzle of whether or not animals perform intentional actions.

Intentional actions are characterized Intentionally and it is commonly

accepted by philosophers that Intentionality presupposes consciousness. 7

What, indeed, could be more obvious? One must be conscious in order to

believe, or in order to want, or in order to perform any other ‘mental act’.

But what is this consciousness that is presupposed? Is it something only people are capable of, or can animals be conscious? If we decide – and this

is fairly common – that only language-users can be conscious, does this suffice to demonstrate that it would be wrong to ascribe Intentionally

characterized actions to dumb animals? Anything that suffices to

demonstrate this must itself be suspect, for we do describe animal

behaviour in Intentional terms quite successfully. 8 The way out of this impasse must be more than a solution to a conceptual problem via analysis of language, for language is deficient in this area. The way out is an

analysis of phenomena at the sub-personal level, and although this leads one into areas many philosophers would prefer to avoid, the alternative is the perpetuation of traditional confusions.

Compared with the ‘common-sense’ vision of awareness and

consciousness, or the literary vision, or even most philosophical visions, the concept of awareness1 is quite austere. It has been laundered of three different sets of connotations that ordinarily accompany the notion of

awareness. First, it detaches awareness from the notion of behavioural

control, and this divorce has been examined and defended in § 15. Second, it allows for no pictorial or imagistic connotations; there has been no talk of

sense-data, images, appearances or any of the other colourful

performers in the ‘theatre of consciousness’. ‘Aware1’ applies to certain hypothetical language-using machines, but in its application to them there is no suggestion that these machines have a rich inner life of psychic

imagery. Since this notion of inner imagery is tenacious, it will be dealt with at length in Chapter 7. Third, I have not left any room in the concept of awareness1 for the sort of creative marshalling of thoughts that is

generally supposed to go on ‘in consciousness’. Events with content just *arrive* at the awareness line, and no mechanism has been suggested that might arrange these, infer from these, consider these, or jump to

conclusions on the basis of these. All one can do with these, to put it crudely, is say them or refrain from saying them. The notion of thinking as an active, creative process in consciousness will be treated in Chapter 8.

7

MENTAL IMAGERY

XVII THE NATURE OF IMAGES AND THE

INTROSPECTIVE TRAP

The view of awareness or consciousness developed in the last two chapters makes it quite clear that we are not aware (in any sense of the word) of mental *pictures*, and although few philosophers these days will express outright allegiance to the doctrine of mental imagery, these ghostly

snapshots have not yet been completely exorcized from current thinking.

Introspection is often held to tell us that consciousness is filled with a variety of peculiar objects and qualities that cannot be accounted for by a purely physical theory of mind, and this chapter is devoted to demolishing this view. The imagistic view of consciousness has been in the past a

prolific source of confusions, such as the perennial problems of

hallucinations, ‘perceptual spaces’ and colour qualities, to name a few.

Once the distinction between the personal and sub-personal level is made clear and mental images are abandoned these problems vanish.

Although the myth of mental imagery is beginning to lose its grip on

thinkers in the field, it is still worth a direct examination and critique. 1 I

shall restrict the examination to visual perception and mental imagery, since the results obtained there can be applied directly to the other sense modalities. We are less inclined to strike up the little band in the brain for auditory perception than we are to set up the movie screen, so if images can be eliminated, mental noises, smells, feels and tastes will go quietly.

The difficulty with mental images has always been that they are not

very much like physical images – paintings and photographs, for example.

The concept of a mental image must always be hedged in a variety of

ways: mental images are in a different space, do not have dimensions, are subjective, are Intentional, or even, in the end, just quasi-images. Once mental images have been so qualified, in what respects are they *like* physical images at all? Paintings and photographs are our exemplary

images, and if mental images are not like them, our use of the word

‘image’ is systematically misleading, regardless of how well entrenched it is in our ordinary way of speaking.

Let me propose an acid test for images. An image is a *representation* of something, but what sets it aside from other representations is that an

image represents something else always in virtue of having at least one quality or characteristic of shape, form or colour in common with what it represents. Images can be in two or three dimensions, can be manufactured or natural, permanent or fleeting, but they must *resemble* what they represent and not merely represent it by playing a role – symbolic,

conventional or functional – in some system. Thus an image of an orange need not be orange (e.g., it could be a black-and-white photograph), but something hard, square and black just cannot be an image of something

soft, round and white. It might be intended as a *symbol* of something soft, round and white, and – given the temper of contemporary art – might even be labelled a *portrait* of something soft, round and white, but it would not be an image. Now I take the important question about mental images to

be: are there elements in perception that represent in virtue of resembling what they represent and hence deserve to be called images?

First let us attack this question from the point of view of a sub-personal account of perception. Consider how images *work*. It is one thing just to be an image – e.g., a reflection in a pool in the wilderness – and another to function as an image, to be taken as an image, to be used as an image. For an image to work as an image there must be a person (or an analogue of a person) to see or observe it, to recognize or ascertain the qualities in virtue of which it is an image of something. Imagine a fool putting a television camera on his car and connecting it to a small receiver under the bonnet so the engine could 'see where it is going'. The madness in this is that

although an image has been provided, no provision has been made for

anyone or anything analogous to a perceiver to watch the image. This

makes it clear that if an image is to function as an element in *perception*, it will have to function as the raw material and not the end product, for if we suppose that the product of the perceptual process is an image, we shall have to design a perceiver-analogue to sit in front of the image and yet another to sit in front of the image which is the end product of perception in the perceiver-analogue and so forth *ad infinitum*. Just as the brain-writing view discussed in Chapter 4 required brain-writing readers, so the image view requires image-watchers; both views merely postpone true

analysis by positing unanalysed man-analogues as functional parts of men.

In fact the last image in the physical process of perception is the image of stimulation on the retina. The process of afferent analysis begins on the surface of the retina and continues up the optic nerve, so that the exact

pattern of stimulation on the retina is ‘lost’ and replaced with information about characteristics of this pattern and eventually about characteristics of the environment. 2 The particular physiological facts about this neural analysis are not directly relevant to the philosophical problem of images.

The nervous system *might* have transmitted the mosaic of stimulation on the retina deep into the brain and then reconstituted the image there, in the manner of television, but in that case the analysis that must occur as the first step in perception would simply be carried out at a deeper anatomical level. Once perceptual analysis has begun there will indeed be elements of the process that can be said to be representations, but only in virtue of being interrelated parts of an essentially arbitrary system (see Chapter 4).

The difference between a neural representation of a square and that of a circle will no more be a difference in the shape of the neural things, than the difference between the *words* ‘ox’ and ‘butterfly’ is that one is heavier and uglier than the other. The upshot of this is that there is no room in the sub-personal explanation of the perceptual process, whatever its details, for images. Let us turn then to the personal level account of mental

imagery to see if it is as compelling, after all, as we often think.

Shorter, in ‘Imagination’, 3 describes imagining as more like depicting –

in words – than like painting a picture. We can, and usually do, imagine things without going into great detail. If I imagine a tall man with a

wooden leg I need not also have imagined him as having hair of a certain colour, dressed in any particular clothes, having or not having a hat. If, on the other hand, I were to draw a picture of this man, I would have to go into details. I can make the picture fuzzy, or in silhouette, but unless something positive is drawn in where the hat should be, obscuring that

area, the man in the picture must either have a hat on or not. As Shorter points out, my not going into details about hair colour in my imagining does not mean that his hair is coloured ‘vague’ in my imagining; his hair is simply not ‘mentioned’ in my imagining at all. This is quite unlike

drawing a picture that is deliberately ambiguous, as one can readily see by first imagining a tall man with a wooden leg and then imagining a tall man with a wooden leg who maybe does and maybe does not have blond hair,

and comparing the results.

If I write down a description of a person it would be absurd for anyone to say that my description cannot fail to mention whether or not the man is wearing a hat. My description can be as brief and undetailed as I like.

Similarly it would be absurd to insist that one’s imagining someone must go into the question of his wearing a hat. It is one thing to imagine a man wearing a hat, another to imagine him not wearing a hat, a third to imagine his head so obscured you can’t tell, and a fourth to imagine him without going into the matter of headgear at all. Imagining is depictional or

descriptional, not pictorial, and is bound only by this one rule borrowed from the rules governing sight: it must be from a point of view – I cannot imagine the inside and outside of a barn at once. 4

A moment’s reflection should convince us that it is not just imagining, however, that is like description in this way; all ‘mental imagery’,

including seeing and hallucinating, is descriptional. Consider the film version of *War and Peace* and Tolstoy’s book; the film version goes into immense detail and in one way cannot possibly be *faithful* to Tolstoy’s words, since the ‘picture painted’ by Tolstoy does not go into the detail the film cannot help but go into (such as the colours of the eyes of each filmed soldier). Yet Tolstoy’s descriptions are remarkably vivid. The point of this is that the end product of perception, what we are aware of when we

perceive something, is more like the written Tolstoy than the film. The writing analogy has its own pitfalls, as we saw in Chapter 4, but is still a

good antidote to the picture analogy. When we perceive something in the environment we are not aware of every fleck of colour all at once, but

rather of the highlights of the scene, an edited commentary on the things of interest.

As soon as images are abandoned even from the personal level account

of perception in favour of a descriptional view of awareness, a number of perennial philosophical puzzles dissolve. Consider the Tiger and his

Stripes. I can dream, imagine or see a striped tiger, but must the tiger I experience have a particular number of stripes? If seeing or imagining is having a mental image, then the image of the tiger *must* – obeying the rules of images in general – reveal a definite number of stripes showing, and one should be able to pin this down with such questions as ‘more than ten?’, ‘less than twenty?’. If, however, seeing or imagining has a

descriptional character, the questions need have no definite answer. Unlike a snapshot of a tiger, a description of a tiger need not go into the number of stripes at all; ‘numerous stripes’ may be all the description says. Of course in the case of actually seeing a tiger, it will often be possible to

corner the tiger and count his stripes, but then one is counting real tiger stripes, not stripes on a mental image. 5

Another familiar puzzle is Wittgenstein’s duck-rabbit, the drawing that looks now like a duck, now like a rabbit. What can possibly be the

difference between seeing it first one way and then the other? The image (on the paper or the retina) does not change, but there can be more than one description of that image. To be aware1 of it first as a rabbit and then as a duck can be just a matter of the content of the signals crossing the awareness line, and this in turn could depend on some weighting effect

occurring in the course of afferent analysis. One says at the personal level

‘First I was aware of it *as* a rabbit, and then *as* a duck’, but if the question is asked ‘What is the difference between the two experiences?’, one can only answer at this level by repeating one’s original remark. To get to other more enlightening answers to the question one must resort to the

sub-personal level, and here the answer will invoke no images beyond the unchanging image on the retina.

Of all the problems that have led philosophers to posit mental imagery, the most tenacious has been the problem of hallucinations, and yet it need hardly be mentioned that there is no problem of hallucinations *unless* one is thinking of awareness imagistically. On the sub-personal level, there can be little doubt that hallucinations are caused by abnormal neuronal

discharges. Stimulation by electrode of micro-areas on the visual cortex produces specific and repeatable hallucinations. 6 Having a visual hallucination is then just being aware1 of the content of a non-veridical visual ‘report’ caused by such a freak discharge. And where is this report, and what space does it exist in? It is in the brain and exists in the space taken up by whatever event it is that has this non-veridical content, just as my description of hallucinations takes up a certain amount of space on

paper. Since spatiality is irrelevant to descriptions, freak descriptions do not require ghostly spaces to exist in. 7

The one familiar philosophical example that may seem at first to resist the descriptional view of perception and awareness in favour of the

imagistic is the distinction, drawn by Descartes, between imagining and conceiving. We can imagine a pentagon or a hexagon, and imagining one

of these is introspectively distinguishable from imagining the other, but we cannot imagine a chiliagon (a thousand-sided figure) in a way that is

introspectively distinct from imagining a 999-sided figure. We can, however, *conceive* of a chiliagon (without trying to imagine one) and this experience is perfectly distinct from conceiving of a 999-sided figure.

From this it might be tempting to argue that whereas conceiving might

well be descriptional and not imagistic, imagining must be imagistic, for our inability to imagine a chiliagon is just like our inability to tell a *picture* of a chiliagon from the *picture* of a 999-sided figure. All this shows, however, is that imagining is like *seeing*, not that imagining is like making pictures. In fact, it shows that imagining is *not* like making pictures, for I certainly *can* make a picture of a chiliagon if I have a great deal of patience and very sharp pencils, and when it is done I can tell it from a picture of a 999-sided figure, but this deliberate, constructive activity is unparalleled by anything I can do when I 'frame mental

images'. Although I can *put together* elements to make a mental 'image'

the result is always bound by a limitation of seeing: I can only imagine what I could see in a glance; differences below the threshold of

discrimination of casual observation cannot be represented in imagination.

The distinction between imagining and conceiving is real enough; it is like the distinction between seeing and listening to someone. Conceiving

depends on the ability to understand words, such as the formula 'regular thousand-sided figure', and what we can describe in words far outstrips what we can see in one gaze.

If seeing is rather like reading a novel at breakneck speed, it is also the case that the novel is written to order at breakneck speed. This allows introspection to lay a trap for us and lead us naturally to the picture theory of seeing. Whenever we examine our own experience of seeing, whenever

we set out to discover what we can say about what we are seeing, we find all the details we think of looking for. When we read a novel, questions can come to mind that are not answered in the book, but when we are

looking at something, as soon as questions come up they are answered

immediately by new information as a result of the inevitable shift in the focus and fixation point of our eyes. The reports of perception are written to order; whatever detail interests us is immediately brought into focus and reported on. When this occurs one is not scanning some stable mental image or sense-datum. One is scanning the outside world – quite literally.

One can no more become interested in a part of one's visual experience

without bringing the relevant information to the fore than one can run

away from one's shadow. For this reason it is tempting to suppose that everything one *can* know about via the eyes is *always* 'present to consciousness' in some stable picture.

To sit and introspect one's visual experience for a while is not to

examine normal sight. When one does this one is tempted to say that it is all very true that there is only a small, central part of the visual field of which one is aware at any moment, and that to describe the whole scene

our eyes, our fixation point, and our 'focus of interest' must scan the sensory presentation, but that the parts we are not scanning at any moment persist or remain, as a sort of vague, coloured background. Of this

background we are only 'semi-aware'. Here, however, introspection runs

into trouble, for as soon as one becomes interested in what is going on outside the beam of the fixation point one immediately becomes aware

(aware1) of the contents of peripheral signals, and this phenomenon is

quite different from the ordinary one. While it is true that one can focus on a spot on the wall and yet direct one's attention to the periphery of one's visual field and come up with reports like 'There is something blue and book-sized on the table to my right; it is vague and blurred and I am not sure it is a book', it cannot be inferred from this that when one is *not* doing this one is still aware of the blue, booklike shape. We are led to such conclusions by the natural operation of our eyes, which is to make a

cursory scanning of the environment whenever it changes and as soon as it changes, and by the operation of short-term memory, which holds the

results of this scanning for a short period of time. In familiar surroundings we do not have to see or pay attention to the objects in their usual places.

If anything had been moved or removed we *would* have noticed, but that does not mean we notice their presence, or even that we had the experience (in any sense) of their presence. We enter a room and we know what

objects are in it, because if it is a familiar room we do not notice that anything is missing and thus it is filled with all the objects we have

noticed or put there in the past. If it is an unfamiliar room we

automatically scan it, picking out the objects that fill it and catch our attention. I may spend an afternoon in a strange room without ever being aware (in any sense) of the colour of the walls, and while it is no doubt true that had the walls been bright red I would have been aware of this, it does not follow that I must have been aware that they were beige, or aware

that they were colourless or vaguely coloured – whatever that might mean. [8]

It is true, of course, that when we see we do not simply see *that* there is a table in front of us, but a table of a particular colour and shape in a particular position and so forth. All this need mean is that the information we receive is vivid and rich in detail. This is not true of the vision of many lower animals. The frog, for example, can see that there is a small moving object before him, but he cannot see that it is a fly or a bit of paper on a string. If the small object is not moving, *he cannot see it at all*, because motion signals are required for the production of the higher-level signals that will initiate a behavioural response. A frog left in a cage with freshly killed (unmoving) flies will starve to death, because it has no equipment for sending the signal: there is a fly (moving or still). Dangle a dead fly on a string and the frog will eat it. [9] The difference in degree of complexity and vividness between frog and human perception does not warrant the

assumption that there is a difference in kind – however much we may feel that a picture is worth a thousand words. 10

XVIII COLOURS

Getting rid of images as things present to consciousness is also getting rid of the qualities that these images would have to have, were there any. For some of these qualities it is a clear case of good riddance. We can all do without the dimensionlessness of mental images (that strange quality that prevents us from putting any kind of a ruler, physical or mental, along the boundaries of mental images), and their penchant for inhabiting a special space of their own, distinct from physical space. With colour, however, there are sure to be misgivings. If anything is a quality, one is inclined to say, colour is a quality, and physics tells us it is not a quality of the ultimate particles that make up the physical universe. Yet colour as a

quality is eminently spatial ('Everything coloured is extended'), so it must exist in the phenomenal space of mental images. The Lockean distinction between primary and secondary qualities has remained compelling in spite of several centuries of rebuttal, and it tends to lead to the view that the primary qualities (the real, physical properties of particles) somehow work to produce in our consciousness the secondary qualities (seen colours,

heard tones, the *feeling* of heat and cold) and that these qualities are on an ontological footing with the primary qualities but somehow insusceptible to analysis and explanation within the physical sciences. They are real, but

'emergent', the essentially novel and unpredictable product of a

fantastically complex collaboration of the primary qualities of particles.

Locke did not, of course, say that primary qualities produced secondary qualities in us, but produced *the ideas of* secondary qualities in us. For Locke, whose view was clearly imagistic, this amounted to much the same thing; an idea of a certain shade of red *was* that shade of red, *had* that shade of red for a quality. But if, in the interests of a truth more important here than any biographical truth about Locke's beliefs, we misread Locke, he can come out saying something that involves no ontological problems

about the status of colours as existing qualities. If ‘idea’ is stripped of its imagistic connotations, having an idea of a colour need no more involve the existence of anything mental that *has* the colour as a quality than having the idea of a unicorn involves the existence of anything that is a unicorn.

The notion that colour words mean what they do in virtue of their being used by people as names for inner, private qualities succumbs to an

argument that has become familiar in many forms: if, say, ‘red’ were the name of a private, inner quality, then one person could not teach the word (its use or meaning) to another, for he would never know whether or not his pupil was associating the word with the right private quality. If the teacher used a collection of red objects as props to help his pupil, he would either have to *assume* that every time he held up a red object his pupil experienced the right private quality, or admit that he could never know whether his pupil had caught on, in spite of the fact that his pupil called all and only those things red his teacher did. The latter alternative is absurd, and the former amounts to a disguised admission that ‘red’ has a *public* reference and criterion; if the pupil’s agreement with his teacher’s public use of ‘red’ is satisfactory to show that the pupil has learned the word, we can, as Wittgenstein said, ‘divide through’ by the private quality, which is superfluous to the analysis. In other words, supposing colour words to

refer to inner qualities is another case of the error, familiar from Chapters

4 and 5, of taking the process to be analysed and using it as an unanalysed part of one’s analysis. In this case, in an effort to understand how a person ascertains the colour of some external object we launch our analysis with the assumption that a person does this by ascertaining the colour of some private, internal object.

If colours are not private qualities, and if they are also not among the primary qualities of physics, we seem almost to be left with the intolerable position that nothing at all is coloured. But this is absurd. When a person says he is looking at something red he is not describing any internal event, he is describing something external, and he can be right or wrong, so

redness must be a property of external objects. But what property? The

facile supposition is that colour properties are reflective capacities of surfaces that can ultimately be characterized in terms of physical

structures, probably at the sub-atomic level, but unfortunately this is already known to be false. The relation between colour experiences (being aware1 that something is red, mauve, green) and light waves striking the retina is not at all the one-to-one correspondence one might expect. A

variety of different combinations of wavelengths can produce the same

'experienced colour', and even monochromatic light can produce, under

certain conditions, the experience of a wide range of colours. 11 Then,

although the sub-atomic characteristics of surfaces that reflect light predominantly of one wavelength can now be described in some detail,

these different types of surface do not correspond neatly to the colours we observe things to be. It is possible, in fact, that two reflective structures both producing the experience of a particular colour might have no

characteristics in common to distinguish them from reflective structures producing different colour experiences except the mere fact that they do produce the same colour experience in people. If this were the case, what would someone be saying when he said something was red? He would be

saying that it had reflective property *x* or *y* or *z* or …, and the disjunction of properties associated with one colour might be very long. 12 If the surfaces seen as one colour did in fact have nothing in common with one another except that they were seen as one colour, should we say that the phenomenon of colour is an illusion, that red and green and blue are not real properties in the world? 13 This can be made plausible by considering the case of colour-blindness.

A man might be colour-blind only to red and green and if he did not

know he was colour-blind, he would probably suppose that grass and ripe apples, fire engines and wine bottles *had something in common* – after all

they were all the same colour (call it *gred*). Since, he might speculate, they are all the same colour, they must share some objective, intrinsic property of their surfaces. This would be an unwarranted leap, however, for the fact that these things were all *one* colour and not two would be due to an idiosyncrasy in the man's visual system, and not any common features of the objects. We might say that the man's experience of grass and ripe

apples as both gred was an illusion. Then what are we to say of normal

colour discrimination? If it happened to be the case that fire engines and ripe apples were not both red 'for the same reason', by parity of reasoning with the case of the colour-blind man should we not say that our seeing them the same colour is an illusion? Suppose there were another race of creatures relative to which human beings were colour-blind. We can

suppose, for example, that they see fire engines as one colour and apples as another entirely different one. If all the surfaces they saw as one colour did in fact have some structural property in common, would we say that

they had *veridical* colour vision, in contrast to human illusory colour vision? Or consider a race with an even more sensitive discrimination

system, one which varied directly with length of light wave entering the

eye. Wouldn't such a creature have the truest colour perception of all, since it would see red if and only if light from the red band of the

spectrum entered the eye? To suppose that either of these races had a *truer* perception of colour than human beings is to lose sight of what we mean by the colour of a thing. The latter race would report that things were constantly changing colour, that you couldn't count on a thing remaining the same colour over even a short period of time, but given what *we* mean by colour, this just is not true and any being that saw things that way would be suffering from colour illusions. The very meaning of our colour words is anchored in such facts as that red things are the ones that look red, and stay looking red, under most conditions. The former race, which saw fire engines as one colour and ripe apples as another, would be

similarly viewed as having *faulty* colour vision, not *truer* colour vision –

for they would be unable to tell when an apple was just the same shade as the fire engine, or when a lady's handbag perfectly matched her coat.

Paintings that we found to have a proper colour balance might seem garish and confused to these creatures; they might see the delicate shadings of a Botticelli as a series of contrasting bands of colours. Where the argument that our colour vision might be illusory goes wrong is in supposing that what we mean when we say something is red or green or blue is that the

thing has some sub-atomic surface structure. In fact what we mean when

we say something is red is just that it is the same colour as ripe apples and glowing embers. This is the true subjectivity of colour qualities: not that they are private, internal qualities, but that red things are all and only those things taken by normal human beings to be red, regardless of their surface structures or reflective capacities. This subjectivity does not, of course, prevent information about colours from being useful. We can rely on our colour vision to tell us that iron is rusting, bananas are ripe, solutions contain copper, snakes are of a poisonous variety. We might

know, for example, that the only snakes in a certain area which were

poisonous had spots of a certain shade of green on their backs, and might hand out samples of this colour on little cards to a group of hikers. What would then be important would not be that the snakes' spots and the paint on the card share some structural property, but only that they be the same colour.

Colour, then, is not a primary physical property like mass, nor is it a complex of primary properties, a structural feature of surfaces. Nor again

is it a private 'phenomenal' quality or an 'emergent' quality of certain internal states. Colours are what might be called functional properties. A thing is red if and only if when it is viewed under normal conditions by normal human observers it looks red to them, which only means: they are demonstrably non-eccentric users of colour words and they say, sincerely,

that the thing looks red. Their saying this does not hinge on their perusal of an internal quality, but on their perception of the object, their becoming aware1 that the thing is red.

8

THINKING AND REASONING

XIX PEOPLE AND PROCESSES

The developing picture of consciousness as merely awareness1 of the

contents of certain states or events leaves no room for the common and

unreflective view of consciousness as the place where thought processes occur. Thinking and reasoning are *things that we do*, not merely experiences of which we are aware, but if thinking is in fact an ‘activity of mind’ or reasoning a ‘process of thought’, we seem to have a dilemma. To repeat Lashley’s dictum, no activity of mind is ever conscious. We have, of course, no introspective access to the actual processes of the brain, but neither do we have introspective access to what may be called the

activities of the mind; we can think a thought in dactylic hexameter, but we have no inkling of how this is done – it just ‘comes to us’ from we

know not where. Yet what is thinking and reasoning if not a ‘conscious

activity of the mind’, an activity of which we are aware?

There are several separable senses of ‘think’. There is a sense related to belief or opinion (‘what do you think of that?’, ‘I think that …’); a sense alluding merely to our ‘stream of consciousness’ (‘I can’t stop thinking about her’); and then there is the sense of interest to us here, connoting purposeful and diligent reasoning, as in the sign on the office wall

‘Think!’. The sign does not exhort the workers to have opinions, nor is it the unnecessary directive to have a stream of consciousness. In some way or other thinking in this sense, or reasoning, is a process, for it takes time, can leave us exhausted, go astray, be difficult, bog down. 1 Yet we are

aware of this apparently internal process in some rather special way, for whereas if I am aware at all of my digestive processes it is only by

observation and inference, my access to my reasoning is more direct.

There are two questions before us, then. What is going on (what process is it) when someone reasons or thinks, and what access does he have to

whatever is going on? Answering these questions will hinge, as before, on drawing the distinction between the personal and sub-personal levels of explanation.

The absurd view of thinking and reasoning, but one which occasionally

infiltrates current thinking about the nature of mind, is that consciousness is an arena into which are led propositions, thoughts, logical operators and

universal rules. The logical operators, like drill majors, direct the propositions into proper marching order, subsuming particulars and

classifying concepts according to the behests of the universal rules, and then produce out of thin air a conclusion to bring up the rear. The audience is the introspective eye which reports this ‘marshalling of thoughts’ to the world at large. Akin to this view is what might be called the ‘hammer and tongs’ view of thinking or reasoning. One supposes that there are

conscious acts of reasoning, acts of judgment and acts using concepts, and on the model of public acts we expect some organ, arm or tool to be *acting on* some object or some raw material – all this within the arena of consciousness. An act, one is tempted to say, cannot be a blank nothing acting on nothing, so there must be both agent and objects in

consciousness. Ridiculing these views is not, however, finding an

alternative that adequately accounts for what is remotely plausible in

them. Something is going on when we reason, and an account of this must be found.

One account that has been offered tries to solve the puzzle by denying

that reasoning is an internal process at all. Reasoning, according to Ryle, is a *social* activity – nothing more than the expounding of already formulated argument, usually to other people: students, colleagues,

juries. 2 This may be one sense of the term, but it is not what one is talking about when one wonders whether animals can reason, or says someone is

not very sharp at reasoning. Were we to adhere to this sense of the term, the question whether animals can reason would receive an abrupt answer: since they cannot expound, they cannot reason. Ryle's attempt to make the public, external presentation of arguments the fundamental and underived activity of reasoning is unconvincing, but it has a germ of truth in it.

Reasoning is not, as Ryle urges, fundamentally a social activity, but it is fundamentally a *personal* activity. As Ryle points out, such quasi-logical verbs as 'conclude', 'deduce', 'judge' and 'subsume' do not refer to

processes at all, but are used in the presentation of results already arrived at. 3 One could not design – let alone construct – a *concluding device* – for the only time one concludes, e.g., that Smith is the culprit, is when one says or writes or thinks to oneself: 'and so I conclude from this evidence that Smith is the culprit', or 'Aha! So it was Smith all along!' One could make a device to utter these words, but just saying the words is not

reasoning. If reasoning is a process, it is not a concluding process, and if it

is made up of operations (e.g., 'logical operations'), none of them will be concluding operations.

The fact that in using such verbs as 'conclude', 'deduce' and so forth we are not describing or naming operations or processes which we somehow

observe does not license the conclusion that there are no temporal

operations or processes going on behind our announcements of

conclusions. People *arrive at* conclusions, and, as the bland verb suggests, this is not a process that people go through or an activity in which they engage, so we cannot ask the question ‘How do you arrive at a

conclusion?’ and expect an answer in the form ‘First I do this, and then I do that’; people do not *do* anything in order to arrive at conclusions, but their brains must. The distinction between the personal and sub-personal levels of explanation is nowhere more important than in the area of

thinking and reasoning. People can reason, but brains cannot, any more

than feet (or whole bodies) can flee or a hand can sign a contract. People can *use* their feet in fleeing or their hands in signing a contract, but it would not be correct to say in the same sense that people use their brains in thinking and reasoning. We say ‘Use your head!’, but if this were

understood in the same sense as ‘Use your index finger!’ said to a

violinist, we would be at a loss to know what to do.

Were we to take what goes on in the brain and analyse it into parts, we should not expect those parts to be, say, concluding or deducing

operations, for that is to confuse levels, 4 and yet some operations of a different sort must occur. When computers are made to perform logical

operations, the abstract, timeless transformations and operations of logic are realized in physical, temporal operations, and the production of results or conclusions takes time and energy. That there must be analogous

processes in the brain can be seen by considering what Bennett calls

‘mental trial-and-error’, the sort of pondering one does when one imagines various outcomes to one’s behaviour before acting. This is a sort of

thinking or reasoning, and ‘sometimes we do this with words, and

sometimes by a kind of imaginative and experimental picturing of the

outcome of various possible courses of action.’5 Saying that reasoning of this sort is a matter of imagining different scenarios cannot be the whole story, however, for how is it that the outcome of a course of action follows in the imagination once we have imagined the course of action?

I can imagine very bizarre outcomes of imaginary actions. I can imagine, for example, picking up a teacup and moulding it in a twinkling into a live rabbit. The fact that in problem solving by pondering these unusual outcomes are excluded from the imagination – even though I may

imagine an outcome which is not right – needs explanation. Imagining a

course of action does not include the outcome automatically if there is anything new or puzzling involved. The mere fact that imagination is

neither a direct transcription or something earlier experienced (in which case it might be stored and then rerun like a film) nor a completely

disconnected sequence of ‘imagery’ must mean that it proceeds in a

regulated way, guided by stored information on experience *in general*.

Although we cannot see or introspect this general knowledge working to

guide our imaginings, we can infer its existence from what we are aware1

of when we imagine.

The fact that we have no introspective access to these internal

operations, whatever they are, is obscured by the fact that we do have

introspective access to *some* operations, or, to put it better: while engaged in problem solving we are aware1 of a series of things prior to arriving at a conclusion, and we can often, on the basis of this awareness1, divide our problem solving into a sequence of *operations* or *steps*. Reaching a conclusion is something that happens, that occurs before or after other things occur, and figuring out the answer to a problem takes time. When one is asked how one figured out the answer, one can often give a list of

steps, e.g., ‘first I divided both sides by two, and then I saw that the left side was a prime …’. What one is doing when one reports these steps is by no means obvious. Are the operations reported in some way *atomic*, or can they be analysed into further operations? Suppose I tell you I first divided eight by two. If you then ask how I did this, I will be left speechless: the operation had no introspectible parts for me, but does that mean it *could* have no further parts, discoverable by some other sort of analysis?

Our inability to analyse introspectively our own problem-solving

activities below a certain level of simplicity strongly suggests an analogy with certain sorts of problem-solving computer programmes. In the field of ‘computer simulation of cognitive processes’, 6 the professed object is to get a computer to solve a problem ‘the same way’ a human being solves it, and typical procedure is to construct one’s computer programme so that

it prints out a play-by-play of its operations in the course of searching for a solution to a problem. This ‘machine trace’ or ‘programme trace’ is then compared with the ‘protocol statements’ of a human subject or subjects

describing their own efforts to solve the same problem. The machine trace of the solution of a particular problem may report many steps –

particularly the mindless trial-and-error sort of series often called ‘brute force’ computing – that do not appear in the subject’s protocol, and in fact are explicitly denied by the subject, e.g., ‘I certainly didn’t methodically check each piece on the chessboard before concluding my rook was

unguarded.’ What can be concluded from these dissimilarities? Some

critics have wanted to conclude that this is evidence that the computer and the subject are using very different methods, or their computations involve different *processes*, but this does not follow. In the case of the computer, there is a certain limit to the depth of analysis of the print-out, determined by the language of the print-out. Ordinarily the print-out is in a high-order language rather than in basic machine language, and hence the computer is unequipped to report the truly atomic steps of its computations, the

opening and closing of ‘logic gates’, for example. Is there a similar limit to the depth of analysis in the human protocol? It is tempting to suppose that when the subject, on introspection, finds that addition of single digits or some other operation is simply unanalysable for him, an atomic

operation lacking introspectible parts, that he has reached the limit of analysis determined by the ‘language’ in which he is programmed for

these particular tasks. It would not follow from this supposition that

addition of digits is, for *human beings*, an unanalysable atomic *process* rather than the complex amalgam of operations it is for a digital computer, but only that it is, for *people*, an unanalysable *activity*: they are not aware1

of any deeper operations. The human print-out capacity in this case might just not go deep enough to reveal the ‘brute force’ computing being done in the brain.

The point becomes clearer when one considers the problem of

‘intuition’. Intuition is often contrasted by the workers in the field of computer simulation to brute force methods of solution, and the

simulators are somewhat in the dark about how one could even begin to

build intuition into a programme. The subject’s protocol to the effect that he just ‘caught on in a flash’ is seen by some to be a stymying indication, but such a protocol could be a case of ‘print-out’ in a language far

removed from the basic operations. A quixotic but illuminating exercise would be to programme a computer to solve certain problems without

providing any print-out capacity except for the standard phrase,

accompanying each solution: ‘It just came to me, that’s all’. Would this not be building intuition into a programme? Intuition, after all, is not a particular method of deduction or induction; to speak of intuition is to deny that one knows how one arrived at the answer, and the truth of this denial is compatible with one’s having arrived at the answer by any

method or process at all, including ‘unconscious’ brute force computing.

Psychologists will never discover a hidden process with the characteristic hallmarks of human intuition, because intuition has no hallmarks.

The analogy between introspective protocol statements and machine

trace print-outs is illuminating, but imperfect. The link between the

internal operations of information processing and human introspection is much looser than that between any computer programme so far developed

and its machine trace. For one thing, as we noticed in Chapter VI, a human introspector can be enticed into speculating, and the line between the two is often hard to discern. When the chess player says something like ‘it was the asymmetry of his bishops that gave me the clue’ he is no longer just accounting what ‘went through his head’, but putting a fallible

interpretation on it. Moreover, the access a person has to the information processing he is doing varies from time to time. Consider the following brace of examples. In case A I walk into the kitchen, pick up an apple and bite into it. When asked why, I remark with surprise ‘Oh! I wasn’t really aware that I had picked up the apple at all. I don’t know why I did.’ In case B I walk into the kitchen, see the apple, say to myself: ‘That is a nice apple I have there, and I won’t spoil my supper, and I like apples, so I think I’ll just pick it up and eat it.’ Here, when I am asked about my

action, I have quite an elaborate protocol to present. But in both cases we can be sure that approximately the same information processing went on, including a lot that did not enter into my protocol in case B. In both cases I would not have picked up the apple had I been in someone else’s house,

nor would I have bitten into a raw egg, nor would I have eaten the apple had I known it was time for dinner. It follows that either the

appropriateness of my behaviour is an immense coincidence or a great

deal of information must have been processed of which I can give no

account in the protocol, e.g., that apples are not poisonous, that it is

socially acceptable to eat apples before dark, and so on virtually *ad infinitum*. This is not to say that all this information need have been processed at this moment, but that earlier processing has prepared me for the appropriate processing I now perform.

Information need not come to the fore, need not cross the awareness

line, for it to contribute to the producing of a conclusion or an inference. If I have stored the information that tomorrow is Friday and I see on the

calendar that on Friday we are dining out, I can say almost immediately that tomorrow we are dining out, without running through the argument in my head. But I cannot do this *just* as soon as I see the calendar; the information from the calendar must be brought to the storage and

operational areas, whatever they may be, which produce the ‘conclusion’

that I can say. Whatever one wants to call these subconscious productions of new information, their operation is essentially logical and they must occur if behaviour control is not sheer magic.

These operations must occur in animals as well as human beings.

Consider the behaviour of certain low-nesting birds that feign a broken wing when a fox appears in order to lead it away from the nest where the unprotected chicks are. The bird’s behaviour may well be only an

‘unreasoning’ tropism, a rigid, inherited routine, but it would not work, and hence would never have become genetically established, if the fox

could not act rationally, unless, of course, the fox’s behaviour is also pure tropism and the entire performance is a stately, ritual dance instinctively performed by hungry predator and alarmed bird, with no benefit accruing to the predator. One might be tempted to adorn the fox’s behaviour and

internal cerebral activity with the postulated ‘mental process’: ‘I like to eat birds, therefore I like to eat limping birds; I cannot catch flying birds, but

this bird is not flying; it is limping, therefore …', but this is silly. In the first place, dumb animals have no language and hence cannot be

aware1 of such thoughts, and in the second place it would be most bizarre even for a person to go through such a tortuous bit of rehearsing to

himself. The thinking of the thoughts, the saying of the words, is not what is necessary, but still the verbal formulae do exhibit, incompletely and vaguely, what must be going on in some internal operations.

In saying these operations are logical, one must be careful not to

suppose that the operations are cases of rigorous, foolproof deduction, governed by the 'laws of logic'. In computers, initial design and

subsequent careful programming can ensure that no operations occur that are not sanctioned by the laws of mathematics and logic, but apparently the organization of the brain is not similarly designed. We can jump to false conclusions and miscalculate arithmetical problems. In particular, there is no need to suppose that the 'logic gates' of a digital computer have their counterparts in the brain, at the 'machine language' level. It is at the gross level of solving problems, plotting trajectories, or generating prime numbers that a computer's operations can simulate, to some degree, the activities of human beings, and the fact that they can do this goes some way towards showing that the operations that make up the gross activities of the computer are similar to the operations 'behind' human problem

solving, but there need not be any binary system, for example,

discoverable in the brain. Our ability to 'follow' the rules of logic in processing information need not be due to any inherited structure ensuring sound, consistent information processing (one thinks here of innate

knowledge of *a priori* principles); we may develop our logical acumen inductively, as part of the development of appropriate afferent-efferent coordination. Part of the way things are is the way things logically are, and if our behaviour is to be appropriate to the way things are, it must be produced along logically sound lines. 7

Should we call this internal information processing reasoning, or thinking, or are there some other phenomena that better fit our intuitions?

If we prefer to heed the ordinary notion that reasoning is a matter of *conscious acts* of the mind, a better way to define reasoning would be as awareness1 of an argument sequence leading to a conclusion. The decision is parallel to the decision on whether ‘aware1’ or ‘aware2’ is *the* notion of awareness. Is introspective access or felicity of behaviour to be the benchmark of reasoning? Consider a mathematician who does a problem in his head without even saying the steps to himself, and when we ask him how he did it, he says ‘I just knew’. Should we say he did the problem without *thinking*? He can tie his shoe without thinking, so why not solve the problem without thinking? Tying his shoe requires some information processing to go on, and so does solving the problem, and if we decide, implausibly, that this is what deserves the name thinking, then, of course, mute animals can think. If, on the other hand, we restrict thinking to something like ‘consciously reasoning with concepts’, then animals cannot think, since they cannot be aware1 of anything, but also people can do many quite intellectual things without thinking.

In the latter sense of ‘think’ we can think enthymematically; in the former sense we cannot. When I say ‘It costs only a pound so it can’t be a real antique’, I leave out many steps in the argument; I do not mention information that must have contributed somehow to the production of my conclusion – information about the shrewdness of antique dealers, the law of supply and demand, the going rates for antiques. If I did not know these contributing facts, if they were not stored in me somehow, I would not

have been able to arrive at this conclusion. It does not follow from this that the logical steps we write down when we present a formal argument

rather than an enthymeme are parallel to distinct operations or events in the brain, but only that the information (including 'supposed' information and misinformation) used in each step must have contributed to the

organization that produced the conclusion. Writing out the logical steps rigorously is thus not being a biographer of any mental or cerebral events, even if the brain does, on a particular occasion, operate rigorously.

Anscombe says, '… if Aristotle's account of the practical syllogism were supposed to describe actual mental processes, it would in general be quite absurd. The interest of the account is that it describes an order which is there whenever actions are done with intentions …'8 An order which is where? It is not an order which there is in our 'conscious thoughts' for we need not think them, and this is what Anscombe must mean by saying the

account of the practical syllogism does not describe 'actual mental

processes'. *Where* the order is is in the Intentional characterization of the brain as an information processor, but this need not be a sequential

ordering of events and operations.

XX REASONS AND CAUSES

We use our reasoning powers not only to solve puzzles but also in what

Aristotle called practical reasoning, to guide and determine our actions. In recounting our reasoning, then, we are not always telling *how* we got a certain solution or conclusion, but often *why* we decided to do whatever we are doing. The *how* and *why* questions can be seen to merge in our ordinary discourse, as when one asks *why* I think my answer is the correct solution to a problem, and I respond by telling him *how* I derived it. This practice of asking and giving one's reasons plays a central role in our notions of action and responsibility, and indeed in our notion of a person; a person performs

actions, and is aware of them and his reasons for them, while bodies (to which sub-personal accounts are appropriate) only

undergo motions. The role of reason-giving will be examined in detail in

Chapter 9; first our capacity to engage in the practice must be examined.

We have seen that often, when a person is asked for a ‘protocol’, his

account drifts imperceptibly away from pure introspection into

speculation, as he is tempted to interpret that of which he was aware1

instead of just recalling it. This tendency produces significant confusions in our notion of reason-giving.

The practice of asking a man for his reasons is accompanied and

explained by the doctrine that a man is the best authority on his own

reasons, and even perhaps a logically insuperable authority. What

accompanies the notion of insuperable authority in turn is the notion of infallible access. Does a man have infallible access to his own reasons?

The infallibility discovered and explained in Chapter 5 was only an infallibility of *expression* of that of which one was aware1, and not at all an infallibility of *detection* of inner processes, events or causes. If reasoning, then, is a process effective in determining our actions, the sort of infallible access described in Chapter 5 will not suffice to give us knowledge of our reasons that is immune to error.

The capacity for awareness1 provides for a sort of knowledge that is

immune to error, which I shall call non-inferential knowledge. One is

aware1 of, and thus knows non-inferentially, what information one is in receipt of – but *not* whether this information is true or false. That is, one

knows non-inferentially that one *seems* to see a man approaching, or *seems* to have a bent knee, for signals (veridical or not) to that effect cross the awareness line. We cannot 'misidentify' the signal (e.g., as signalling an elbow itch rather than a bent knee), but we can go on to interpret the signal as veridical, and then room for more than merely verbal error is

introduced. The case of pain is interesting in that a report of pain has, as it were, a built-in 'seems-to' operator. When one is aware1 that one has a pain in the foot, the signal to that effect cannot be misidentified and *amounts to* having a pain in the foot. If it is veridical then one has an injury in one's foot, but if it is not veridical one still has the pain. To have a pain is to seem to have an injury, so the idiom 'I seem to have a pain'

contains a redundant and meaningless disclaimer; it amounts to saying 'I seem to seem to have an injury'.

When one has non-inferential knowledge of a pain, or of seeming to see

a man approaching, one can have inferential knowledge of an injury, or of a man approaching. In some cases the inference is a conscious one. That is, one is first aware1 that one *seems* to see a man, and then is aware1 that *since* one seems to see a man most likely there *is* a man. This happens only in the most unusual circumstances – when, for instance, one is

expecting an optical illusion, or suspicious that one may be hallucinating, or just engaging in a thought experiment for philosophical ends. More

usually the inference is subconscious, a *fait accompli* that involves no *thinking* (one is aware1 of no argument). The distinction, then, is logical; it distinguishes one evidential status from another. It should not be confused with a psychological distinction between inferences we happen to have

made consciously, and things we know (regardless of evidential status)

without having made any conscious inference. Inferential knowledge is

knowledge where there is logical room for an inference, and hence room

for more than just verbal error.

Now when one responds to a question about his reasons, with what sort

of knowledge does he answer? If one is asked merely for the ideas that are passing through his mind, one can respond in a foolproof way, with only verbal slips to worry about. But when one is asked to give one's reasons for an act, one is asked to give the reasons that 'actually worked', that led to or determined the act, and not just any plausible reasons that come to mind. There is a genuine ambiguity in the demands we make of people

when we ask them to give their reasons. On the one hand we grant

authority to the actor; if he says his reason for taking a drink is that it will calm his nerves, and we believe he is sincere, then that *is* the reason he took a drink. The psychoanalyst, however, may tell us that the *real* reason he took the drink was because he thinks drinking makes him masculine.

What is the *real* reason? If the reason the psychoanalyst gives us is the real reason, this might often (or even always) be a reason of which the actor has no inkling, but if this is what we are asking for when we ask for reasons, why do we ask the actor? His best response, under these 'rules' of the language game, would be to say 'I haven't any idea; you had better ask my analyst.' Or, if he is a fan of psychoanalysis, he may try to

psychoanalyse himself, and say 'Let's see. I wouldn't be at all surprised if my reason for taking the drink is that I have a death wish and am trying to drown myself.' Surely this would be an inappropriate response to our

question; we are not asking a person to psychoanalyse himself when we

ask him for his reasons. On the other hand, we are not merely asking him to report his thoughts to us. Suppose I am asked why I took a drink and I think back and realize that I had just thought: 'I am very upset and a drink would calm my nerves, so I'll have a drink', I may respond by announcing that I wanted to calm my nerves, and this may or may not be a *sincere* answer. If I know perfectly well that I am a compulsive drinker always

looking for a good excuse for a medicinal nip, I would not believe I was reporting my real reasons in saying this. If I am not strong on self-knowledge, on the other hand, I may report this as my reason quite

sincerely. The fact that the thought ran through my head may be

undeniable; that it veridically represents the ratiocination that determined my behaviour is a matter of interpretation and fallible inference.

If without stopping to think I pull a child away from a fire and am asked to give my reasons, I may say with a high degree of conviction, 'because I saw he would soon be burned', and in this case I am not relying on any

remembered thoughts that may have run through my head – for none did –

and my knowledge that this is the correct reason is based on my

knowledge that I did *not* think 'I'm going to kidnap that child' or 'let's put baby in his crib for a while'. Lacking any evidence for exotic explanations, I infer (usually subconsciously) that I recognized the danger and then

acted on this recognition – which is the obvious explanation.

Of course, even when I rely on remembered thoughts in giving my

answer, there is a mediation which gives rise to the possibility of error –

just because I am remembering what I *was* aware1 of, not expressing what I *am* aware1 of. Any memory can be false, of course, and though I cannot be mistaken in thinking that I *seem* to remember being aware1 of a particular thought, I may be misremembering. When we ask a person to

'think out loud' while solving a problem, the protocol we get from him

will thus be more reliable than the one we could get if after finding the solution he was asked to recall the steps he took, just because it eliminates one layer of inferential knowledge; it does not depend on the faithfulness of the reports of his memory. Thus in most cases of giving my reasons my report is doubly inferential: I infer that my memories of my conscious

pondering are sound, and moreover that that pondering was not mere

rationalizing.

The question whether a reason is a bit of rationalizing or truly *my* reason for acting is the question whether I acted *because of* this reason, and this

'because' is a causal 'because', not a reason-giving 'because'. 9 I may have a reason for doing X, may do X, but not because I had the reason (see p.

39); when I *do* X because of the reason, it is not that my reason for doing X is that I have this reason, for that would lead to an infinite regress. The ordinary practice of asking for reasons is predicated on the *assumption* that our conscious reasoning is a reliable manifestation of the information processing that determines our actions, and psychoanalysis is predicated on the counterclaim that it is not. Our ordinary granting of authority to the actor's reports is well supported because in so many instances it just is not to the point what one's deepest source of direction is: when I ask why you are sawing the plank and you tell me you are making a table it is virtually inconceivable that you are mistaken and quite irrelevant that you may

have deep and terrible reasons for making a table, unknown even to

yourself. On the other hand, our willingness to grant authority

occasionally to the psychoanalyst is well supported because we have seen many times that a person's apparently sincere reports of reasons do not harmonize with his behaviour: for example, the proven success of

advertising campaigns based on 'sex appeal' shows quite clearly that we do buy products for reasons other than the hard-headed practical reasons we sincerely avow.

In some cases I may say to my interlocutor that I do not know for what

reasons I did something, and in these cases there still may *be* reasons for

what I did. In other cases I may reply that there was *no* reason for what I did, and I can be right or wrong about this. I may believe that what I did

was sheer doodling, or pointless motion, but this motion *might* have some

'deeper significance', *might* have a reason. In some cases of bodily motion we know enough about the actual mechanisms of control to say with a high degree of certainty that there was no reason, as in the case of the reflex kick. In these cases we are apt to say that the behaviour was *caused* but had no reason. The same verdict is often reached in more speculative

cases. Consider Anscombe's example: ' "Why did you jump back suddenly

like that?" "The leap and loud bark of that crocodile made me jump".'[10]

What sort of knowledge do we have of this sort of 'mental cause'? Not

non-inferential knowledge. It is true that there is something I know about my jump that another person cannot (at this time) know, and so in a way I have 'privileged access' to such mental causes. I know that I saw (or

seemed to see) the crocodile leap and bark, and I know (by 'proprioceptive feedback' from joints and muscles) that I jumped (or seemed to jump),

and *so far as I know* nothing else entered into the situation. What I can know that another person cannot is what was *missing* from the experience.

I know I did not think to myself, just before the crocodile barked, 'I think I'll just jump back for the fun of it', and I know I am not afflicted with some malady that makes me jump every now and then. So I *conclude*

(consciously or subconsciously) that it was the sight of the barking,

leaping crocodile that made me jump, and this conclusion is fairly safe.

But I do not have non-inferential or immediate knowledge of the cause of my jump, and of course it is only contingent that another person cannot know what I do about my jump. Neurologists might someday know just as

well – in fact better – what caused my jump. I have no access, private or otherwise, to my cerebral processes, but only to my awareness and the

succession of messages arriving there; having no other explanations of the jump, and having seen others jump when presented with sudden, strange

sights, I infer (consciously or subconsciously) that the startling sight caused the jump.

Suppose I am crying, and someone asks why. I say ‘because Smith just

died’, and this is a cause, not a reason, for my crying, for I am not crying on purpose or deliberately. In making this report I am assuming, again, that there is a causal relation between learning the sad news and crying, since the occasion is similar to other occasions on which people have

cried. The regularity with which the receipt of sad news is followed by crying suggests that there is a causal relation between the two, and

neurologists may someday provide detailed confirmation of this

hypothesis. But all that I, the crier, may know that another person may not, is that in *this* case nothing else of conceivable relevance, such as an onion or directions in the script: ‘cry here’, has entered into the case.

The fact that such knowledge of causes and reasons is inferential is

obscured when one looks only at the personal level account of what is

going on. I, the person, do not make an inference (consciously), so it is tempting to say that I *just know* what these reasons or causes are, and then the case is easily confused with the cases of genuine experiential certainty, such as my infallible, non-inferential knowledge that I am in pain, which have a different evidential status altogether. Perhaps in cases of inferential knowledge we should not say that *I* make the inference, since it is made subconsciously, but it *is* made, and this is enough to give *my* knowledge of these reasons and causes a mediated evidential status.

Before leaving reasons and causes I want to dispose of a common

misunderstanding to the effect that where there are reasons there are no causes, and vice versa. Anscombe, for one, if I understand her, does not

wish to speak of an action being caused if it occurred for a reason. She says:

… how would one distinguish between cause and reason in such a case as having hung one's hat on a peg because one's host said 'Hang up your hat on that peg'? … Roughly speaking – if one were forced to go on with the distinction – the more the action is described as a mere response, the more inclined one would be to the word 'cause'; while the more it is described as a response to something as *having a significance* that is dwelt on by the agent in his account, or as a response surrounded with thoughts and questions, the more inclined one would be to use the word 'reason'. But in very many cases the distinction would have no point. [11]

There is, of course, causation in both cases – and reasoning in both cases. Reasoning, or its subconscious counterpart, must be going on even when one 'unthinkingly' hangs one's hat on the peg; the behaviour is appropriate to the stimulation because it is mediated by organizations established by stored information – about manners, pegs, hats, and so forth. The 'unthinking' response is leagues beyond the Pavlovian conditioned response (people are not trained to hang their hats up on pegs upon hearing verbal cues), and leagues more beyond the knee-jerk, which is genetically wired in. How very strange it would be that a person should hang up his hat in response to a verbal cue, unless it were a swiftly *reasoned* response.

Support for the view that what we do for a reason does not have a cause is often found in the claim that what we do for a reason we do

intentionally, and *part of what we mean* when we say that an action is intentional is that it is not caused. This claim refers us to an ‘ordinary’

question and answer sequence:

(1) Were you caused to do that (e.g., spill the coffee)?

(2) No, I did it intentionally.

I doubt that this is ordinary. I doubt that anyone would ever speak just this way, but even if it is granted that we do speak this way, what of it? How do I *know* I was not caused to spill the coffee? Have I non-inferential knowledge that I was *not* caused to do it? Would it not be better to say that I have inferential knowledge that at least certain *sorts* of causes were absent? That is, I know I did not feel anyone bump my arm, I know that I am not an epileptic, and I know moreover that I just had the malicious

thought: ‘Let’s make a mess of Smith’s carpet.’ Question (1) asks if

anything like a bump or a twitch or a startling sight caused me to spill the coffee and, as far as I know, nothing like that did cause me to do it. It would be absurd to suppose that when one asked (1) he intended to cover all physical and metaphysical eventualities with regard to causes, and that (2) is anything like a firm assertion of the absence of causes. Thus (1) is, plausibly, an ellipsis for

(3) Did you do that because some external object or internal

malfunction moved your body?

and (2) is an ellipsis for

(4) No cause of *that* sort operated – to the best of my knowledge. I did it intentionally (and I really have no idea what sort of causes if

any *that* might involve).

Thus not only is it the case that when I do something for a reason, what I do is caused, but what makes a reason my *real* reason for doing something is that the events of information processing which cause what I do have among them an event with the content of my real reason, whether or not I am aware1 of this content.

9

ACTIONS AND INTENTIONS

XXI INTENTIONAL ACTIONS

The concept of personal action is an essential adjunct to the concept of a person, for, as we have seen, it is only on the personal level that

explanations proceed in terms of the needs, desires, intentions and beliefs of an *actor* in the environment. Beyond this, the concept of action plays a critical role in our notions of responsibility and punishment. It is well worth detailed elucidation, therefore, and all the more so because once again the traditional views of intentional action will be seen to founder on a failure to make clear the distinction between the personal and sub-personal levels of explanation. The first step is to characterize the class of intentional actions, and since this task has been brilliantly executed by Miss Anscombe in *Intention*, I can do no better than to give a *précis* of her analysis, making a few alterations along the way and then wedding the

results to our emerging picture of awareness1.

First, she points out, it is not bodily motions, but motions under

particular descriptions that are intentional or unintentional. I may be sawing a plank, and it may be one of Smith’s oak planks, so that I am

sawing a plank, sawing one of Smith’s planks, and sawing an oak plank.

These are not, however, different motions; in doing all three at once I am not performing three separate feats of motion. The *action* of sawing the plank (the motions considered under *that* description) can be intentional,

while the *action* of sawing one of Smith's planks (the same motions under a different description) is not.

A necessary condition for membership in the class of intentional actions is that the actor be aware of the action, i.e., aware of the motion under a particular description. If I am typing, and someone asks me 'Why are you tapping out the rhythm of "Rule, Britannia"?' and I reply 'Oh, am I doing that? I was not aware of it', I show (if I am truthful) that the action of tapping out the rhythm was unintentional. I am aware of the motions of

typing, but not under that description, and so, not being aware of tapping out the rhythm, I cannot be doing it intentionally. Hence, intentional

actions exhibit Intentionality, or, more circumspectly, intentional action ascriptions are Intentional contexts.

Awareness, however, is not a sufficient condition. I may be aware that I am doing one thing in the course of doing something else, and yet not be doing the former intentionally. I may notice that I happen to be neatly not stepping on the cracks of the pavement, and yet not be intentionally

missing them; they just happen to match my normal stride. I may, in fact, pay particular attention to this phenomenon to see how long I can keep to my normal stride before my foot lands on a crack, and in such a case I am still not missing the cracks intentionally, but only keeping to my normal stride intentionally – however much I *want* to keep on missing the cracks.

Anscombe takes account of this by distinguishing intentional actions as members of a subclass of the class of actions of which one is aware: the class of one's actions of which one is aware *without observation*. That is, one denies that an action is intentional if one says 'I only *observed* that I was doing that'. 1 When I say I only observed that I was doing something, I mean I saw with my eyes, or heard with my ears, or felt that I was doing something (e.g., inadvertently making scratches on the table with my pen), or even that I knew proprioceptively what I was doing: I knew my knee

jerked without looking, because I had the kinaesthetic sensation of a

jerking knee. If these are the *only* ways I know or am aware that I am doing X, then doing X is not intentional. In what *other* way, then, could I have knowledge or awareness of my action, so that it could be intentional?

The question can be reformulated in the terms developed in previous

chapters for explaining awareness: an action falls into the class of actions

'known without observation' if a signal with a content descriptive of the action crosses the awareness line, and this signal is neither a

proprioceptive signal from muscles or joints nor a signal mediately from the sense organs. Now could any signals fulfil these conditions?

It was argued earlier that last-rank motor signals can be ascribed only uninteresting content, such as 'contract, muscle! ', but that the higher-level directing efferents would, if they could be pinned down at all by the investigator, be amenable to more interesting content ascriptions, viz., *commands* to perform *actions*. For example, 'open the door!' could be the content of a relatively high-level efferent event or state which controlled a number of different sub-routines all 'designed' to get the door open in one way or another. Then if such an efferent command were to send a signal

also across the awareness line, one would be aware1 of its content; one would be aware1 that one was trying to do X, and one's awareness1 would

be of a *particular* description under which the resulting motions were to be subsumed. The sort of aware-ness1 one had of intentional actions would then differ from other sorts of awareness1 in having an *efferent* rather than *afferent* source. Then knowledge without observation can be construed as knowledge of efferent controls, not knowledge of afferent input. That is, one would be aware1 that one was opening the door (or trying to) and not via any afferent route, and hence one could say what he was doing.

Anscombe says several things to support this characterization of our

knowledge of intentional actions. This knowledge, earlier characterized only as non-observational, becomes 'practical knowledge', and is

compared with the knowledge of a man ‘directing a project, like the

erection of a building which he cannot see and does not get reports on, purely by giving orders.’2 The knowledge we have in these cases is knowledge of the *orders* we give, and our beliefs about the state of the building at any time will be correct, provided our orders are carried out.

As Anscombe says, if they are not, ‘Theophrastus’ remark holds good:

“the mistake is in the performance, not in the judgment”.’3

Suppose a man with a hammer decides to finish driving a nail into a

door. Suppose that he happens to see the nail sticking out, and this visual input has the effect, after running through various cerebral organizations, of giving rise to an efferent state or event with the content ‘drive in the nail’. (Note that no appeal is made to anything like ‘conscious decision’.) Then suppose this efferent state or event sends a signal across the

awareness line; the man will then be aware1 of *what he is about*. He may say to himself or out loud ‘about time I finished driving in that nail’ – or he may only be aware1 of what he is about and never go through the longer temporal process of formulating an utterance aloud or to himself. To vary the case, suppose the man is prompted to make a dent in the wood next to the nail – for some arcane reason, or just to be silly. A *different* efferent state would direct this activity and the man would be aware1 of this

different action were the efferent state to send a signal across the

awareness line. Then suppose in the former case the man accidentally

misses the nail and makes a dent; the muscular and skeletal motions might be quite indistinguishable in the two cases, but they would be controlled by different efferent states, and the man would be able to tell the

difference – not because he recognized a qualitative difference in efferent

states (he would have no inkling of what his efferent states were) but because these different efferent states produced in him different

dispositions to express certain things, produced in him different contents of awareness.

How do we know when we have made a mistake or through accident

failed to achieve what we are trying to achieve? Anscombe gives an

example where the utterance of an action description and the action do not match:

…I say to myself ‘Now I press Button A’ – pressing Button B – a

thing which can certainly happen … And here, to use

Theophrastus’ expression again, the mistake is not one of

judgment but of performance. This is, we do *not* say: What you

said was a mistake, because it was supposed to describe what you did and did not describe it, but: What you *did* was a mistake,

because it was not in accordance with what you said.4

As she points out, this is just like obeying an order *wrong*, which is not a case of disobedience, but of malfunction. There are a number of places

where the malfunction can occur, however. It can occur in the

implementing of the button-pressing action or of the verbal utterance – I may have made a slip of the tongue, and have *meant* to say ‘Now I press Button B’. In the latter case I can correct my utterance via ‘feedback

loops’. There is still one more place for malfunction to occur, however.

Malfunction could occur between the directing efferent state and the

awareness line, so that I would be aware1 of ‘ordering myself’ to push

Button A while the efferent state actually directing my motions had the content ‘press Button B’. This would be a case similar to those discussed in § XX, where my sincere report of what I am about is not veridical,

where my awareness1 is not a reliable manifestation of my inner direction.

In this case I cannot be mistaken about that of which I am aware1 – for only correctible verbal slips are possible there – but only mistaken in supposing that I have practical *knowledge* of what I am doing; i.e., in supposing that the content of my awareness1 is a *veridical* report of my inner direction.

The possibility of this distinct source of knowledge is easy to overlook in the course of a casual examination of one’s own experiences just

because ‘practical’ knowledge of our overt actions never exists unmixed.

There is a contingent, functional interdependence of ‘practical’ and

proprioceptive knowledge, so that we are constitutionally unable to write words on a blackboard while blindfolded (to use one of Anscombe’s

examples) without being informed proprioceptively, and so we can become aware1 of the proprioceptive information. Or, without becoming aware1 of the proprioceptive information, we can still be informed by it; i.e., it can contribute to our (inferential) knowledge that not only are we ‘ordering’ a certain action, but it is being correctly carried out. For some actions, such as drawing a cow, visual information is required in addition to

proprioceptive information if I am to know what I am doing (unlike the

action of, say, writing one’s signature with one’s finger in the air), and in these cases if I am blindfolded I will only know what I am trying to do, not what I am doing. It is often seen as a puzzle that a man can quite directly (without conscious inference or observation) know what he is doing, and not merely what he is trying to do, but this puzzle dissolves when one

recognizes that the *experiential* directness (lack of conscious inference) masks an epistemic mediacy, in which information from two or more

sources combines to give us the contents of awareness. For overt actions, involving skeletal and muscular motions, it is contingently impossible to isolate cases in our normal experience, in which we have only practical knowledge – knowledge of the efferent 'commands' given. We are all,

however, familiar with the experience of *trying* to move an arm that is

'asleep', and in these cases our knowledge of what we are trying to do is pure practical knowledge, purely a matter of being aware1 that this is the efferent command being given.

There are thus a variety of ways in which we can be informed about

what we are about. We can *see* what we are doing just the way we see what others are doing; we can feel, proprioceptively, what we are doing; and we can have practical knowledge of what we are doing, that is, we can be

aware1 of our efferent commands. The latter mode of knowledge is a

necessary condition of intentional action, but it is still not a sufficient condition. One may have practical knowledge of what one is doing and

still not be doing it intentionally if one's answer to the question 'Why are you x-ing?' is 'no particular reason' or 'I don't know; I was just doodling'.

As Anscombe points out, such an answer is not a rejection of the question, as 'I was not aware I was doing that' is. 'The question is not refused

application because the answer to it says that there is *no* reason, any more than the question how much money I have in my pocket is refused

application by the answer "None".'[5] The final requirement for an action to be intentional is that there must be a reason that can be given by the actor for the action. Where there is no reason in the offing, as in doodling, Anscombe would call the action voluntary, but not intentional, and this seems harmonious with our ordinary usage. Other actions may qualify as

voluntary, but at least all actions of which one has practical knowledge but can offer no reasons for doing are voluntary. In the case of doodling, for

example, the psychoanalyst may claim – even correctly – that there is in fact an unavowed deeper reason for the doodling, but this does not make the action intentional. This point will become important in § XXIII, when we consider the significance of intentional actions, for in finding a reason for the apparently inadvertent and unintentional, is the psychoanalyst not claiming to put the action on a par with intentional actions? That is, should we not *treat* the action as if it were intentional?

The giving of reasons by the actor is not a foolproof activity, as we saw in § XX. Even where conscious reasoning occurs, so that we are aware1 of the apparent input and output of the reasoning process – though not of the process itself – we cannot know with certainty that what we offer is

reasoning and not rationalization. When conscious reasoning has not

occurred our reason giving is simply conjecture, although often highly

reliable conjecture. The ‘Why?’ routine brings this out: ‘Why are you

sawing the plank?’, ‘I’m making a table’, ‘Why are you making a table?’,

‘To put our food on’, ‘Why put your food on a table?’, ‘Just because, that’s why’. After the initial answer or two, what follows is conjecture or

fabrication, not based on any conscious reasoning one remembers having

performed. It is true or false, however, since either the information cited has contributed indirectly to one’s behaviour or not. The repetition of the

‘Why?’ question is supposed to have the effect of probing deeper and

deeper into the beliefs and methods of the actor, but once the responder has reported whatever thoughts he was initially aware1 of that might have contributed to the direction of his behaviour, the subsequent answers are merely parts of his own personal theory of motivation. It still makes sense to ask the question, though, for the responder has his memories of past

behaviour and the thinking that accompanied it, and this is certainly relevant information not held by the questioner. The actor is simply

empirically better acquainted with his own style of behaviour than anyone else is, although he may not be particularly perceptive or critical about his own behaviour.

The distinction between awareness1 of efferent commands and

awareness1 of one's further reasons is not a sharp one. An intentional

action is a motion under a particular description, and we saw that sawing the plank and sawing one of Smith's oak planks are two different actions.

For an action to be intentional, moreover, we saw that the actor must be aware1 of the action under that description, via efferent commands. Yet is it reasonable to suppose that different *motor* commands would have the different contents 'saw the plank' and 'saw Smith's plank'? The

differentiation here must come at the level of awareness of reasons, not awareness of motor commands, but it is not always clear where to draw

this line. Does one push the pen *in order to* sign the document or *in signing* the document does one push the pen; does one sign the document *in order to* close the deal, or *in closing* the deal does one sign the document? One can say that one must be aware of one's further reasons for performing an intentional action, or one can say that one must be aware of a wider description of what one is doing when one performs an intentional action. The demand for reasons for intentional actions is not a demand

with fixed limits, since there is no fixed length for the nested reasons one must give, and no fixed description from which to start. Sawing the plank is intentional if the question 'Why are you sawing the plank?' is answered

'I'm making a table', but if someone a bit more observant asks, 'Why are you making a table?', this requires something in the way of further reasons if it is to be considered intentional. [6]

XXII WILLING

The account of intention that has been given includes no talk about

volitions or willing. That is because, as Anscombe argues, the verb ‘to will’ is a hoax. There are no such things as acts of will or volitions.

People sometimes say that one can get one’s arm to move by an act

of will but not a matchbox; but if they mean ‘Will a matchbox to

move and it won’t’, the answer is ‘If I will my arm to move in that

way, it won’t’, and if they mean ‘I can move my arm but not the

matchbox’ the answer is that I can move the matchbox – nothing

easier.[7]

The idea that willing is some sort of radiation generated by gritting the teeth and saying, ‘move, move, move’ is hopeless. It arises, no doubt,

from such experiences as lying in bed and saying to oneself ‘I must get up, I must get up; it’s late. On the count of three: one, two, three …’ until finally one gets up. The causal link in these cases has been debated at great length, for on the one hand thinking these thoughts often seems to help or even cause the action, and yet on the other hand very often thinking the thoughts has no effect at all.

It is supposed, perhaps, that when thinking these thoughts does not

work, one is just not thinking hard enough or with enough conviction, but these explanations are obvious dead ends. The ‘tone of voice’ with which one says these things to oneself clearly does not make any difference, and what else can one do to simulate or bring about conviction? The facts of the matter – that sometimes the thoughts seem to help and sometimes not

– suggest that thinking to oneself is merely an accompaniment or by-

product of the actual business of determining action. It seems most likely, that is, that if I have the conviction that I must get up, I can say the words and will thereupon arise from bed – but I will arise equally well if I do not say the words to myself. And if I do not have the conviction that I must get

up, I will not get up, whether or not I mutter exhortations to myself with great vehemence.

The view of neural activity so far developed provides a plausible if

sketchy explanation of this phenomenon. Roughly, in order for the brain to

initiate the activity of getting up, its input must be such that it outweighs, say, the pleasure of just lying in bed, the influence of stored information to the effect that getting out of bed is unpleasant, the input to the effect that the body is still tired, and so forth. As soon as the balance is tipped, the brain initiates the activity and one gets up. There is no need to suppose that once the balance is tipped something must recognize that the balance is tipped and then proceed to *will* the act of getting up, any more than a spinning top must recognize that its gyrostatic force is no longer strong enough to balance it in order to 'decide' to fall over.

When the balance has been tipped or is being tipped in the brain,

messages that cross the awareness line to the effect that it is time to get up, there is much to be done, and so forth, are merely accompaniments to the 'decision-making' of the brain. But when the balance has not been

tipped, no amount of repetition of these messages, if in themselves they are not enough to tip the balance, will bring about the action.

The way, presumably, to tip the balance is to increase and facilitate the sort of input that would outweigh one's inertia. The input that would

accomplish this would depend on the dominant organizations at the time.

The information: 'the British are coming!' would seldom serve to tip the balance. The information 'it's time to get up' can tip the balance if the person has some reason to get up on time or some natural tendency to

regularity and punctuality, which amounts to a 'weakness' for this sort of input. But if the information 'it's time to get up' is not sufficient to outweigh the inertia, no amount of feeding this information repeatedly

into the relevant parts of the brain will tip the balance. It is not awareness or consciousness, however, that is producing this information, but the rest of the brain; awareness is simply the verbal outlet for the information.

Just as one person can devise input information to stimulate another

person – a stubborn or lethargic person – to do one thing and not another, so perhaps the brain, when more or less stalled, resorts to self-stimulation by the production or retrieval of information. And the success in both

cases, of course, would depend on the relevance and abundance of the

information produced. Something like this may well go on, but if it does, awareness of the information, or saying the information to oneself with feeling, would add nothing to the process. As suggested in Chapter 6, it may be that any information that is boosted into the higher levels of neural organization must come to awareness as a matter of physical fact, but then

it is still the boosting and not the awareness itself that is necessary. The notion that must be avoided is that awareness is in any way a centre *from* which efficacious signals, volitions, or any sort of psychic radiation

emanates.

Anscombe touches on this, but in quite different terms.

We can imagine an intention which is a purely interior matter

nevertheless changing the whole character of certain things. A

contemptuous thought might enter a man's mind so that he meant

his polite and affectionate behaviour to someone on a particular

occasion only ironically, without there being any outward sign of

this (for perhaps he did not venture to give any outward sign) …

Let us suppose that the thought in his mind is 'you silly little twit!'

Now here too, it is not enough that these words should occur to

him. He has to mean them. This shows, once more, that you cannot

take any performance (even an interior performance) as itself an

act of intention.8

There is an internal difference, quite clearly, between just saying ‘you silly little twit!’ and meaning it, but this difference is not itself a performance.

The difference must depend on what the function is of the part of the brain that produces this message in awareness. If the message is produced in the course of the brain’s maintaining a particular antagonistic state, if the production of this message is caused by some neural activity that, say, brings into play stored information on the shortcomings of the ‘silly little twit’, then the ‘thought is meant’. If, on the other hand, the message is produced in the course of, say, mere experimentation, such as seeing if it is in fact possible to be polite while thinking ‘you silly little twit’ and not meaning it, then the ‘thought is not meant’. Suppose the thinker of this phrase has been trying to think of a four-word phrase with internal

assonance; suppose in other words someone has just said ‘give me a four-word phrase’ with as much assonance as ‘Philip spilt the milk’, or for one reason or another this task has just occurred to the person. The neural mechanism that produces the message ‘you silly little twit’, like the neural mechanism that produced in me ‘Philip spilt the milk’, has, or can have, virtually no other effect on behaviour or neural state than the production

of words. The activity involved does not influence or mesh with any other activity.

How does one know whether one means it or not? One knows this

simply because one knows what one is about; and one knows this *only* by knowing what messages preceded and followed the message in question.

Imagine a person all of a sudden finding himself saying to himself ‘you silly little twit’. What if no other rancorous thoughts had been going

through his head; what if there was no obvious candidate for the epithet; what if further the thinker had not just been aware of thinking he would try this little experiment? Could there be anything intrinsic in the mere

unheralded, unaccompanied phrase occurring in his awareness that would

tell him whether or not he meant it? Strange, isolated thoughts do spring to people’s minds occasionally, and they can be totally baffled as to the

meaning or importance of these thoughts.

The point that emerges is that awareness is not the home or origin of

intentions or volitions. In fact we have only limited and fallible access to the mechanisms that direct our behaviour. Nothing that goes on ‘in

awareness’ can be construed as an act of will or volition, and nothing that is subconscious would fit the ordinary connotations of these words. Once again, getting rid of the little man in the brain, this time in the guise of

‘conscious agent and source of volitions’, also involves getting rid of the tools of his trade.

XXIII THE IMPORTANCE OF INTENTIONAL ACTIONS

The class of intentional actions has now been characterized as the class of motions under particular descriptions of which the actor has practical

knowledge and for which he is prepared to offer reasons. It is not at all clear why these conditions should make intentional actions so special. The concept of intentional action is critical in our conceptual scheme, for our bestowing and withholding of praise and blame is generally tied to the

decisions we reach regarding intentions. Quite literally a lot hangs on our ascriptions of intentions, and we cannot answer the question of their

importance by pointing out that a lot hangs on them; it is just this that is in need of explanation and eventually moral justification. Unless there is some important, efficacious difference between those motions that are

intentional actions and those that are not, the distinction is pernicious.

A brief look at the range of bodily motions and actions shows that there is no clear-cut line between the intentional and unintentional or the

voluntary and involuntary from the point of view of causal determination.

The circuitry that causes these motions varies only in complexity and

degree of mediation. The only ‘indeterminacy’ that can be held out for the causal sequences governing intentional or voluntary action is due to our lack of knowledge of the nervous system, not to any random effects. 9

Moreover it is now widely recognized that causal indeterminacy is not the

‘freedom’ we should look for to account for ‘freedom of the will’. Nor

does mere complexity of causal antecedents promise to be the important

distinguishing characteristic. As we saw in Chapter 8, even in cases of casual acts where there is no conscious reasoning and no awareness, such as picking up and biting into an apple, the complexity of the causal

antecedents, characterized either extensionally or Intentionally, will be roughly equal to their complexity in cases of intentional action. These are considerations of the wrong sort; assigning responsibility for actions

cannot hinge on the complexity or determinacy of causal sequences, for

when one speaks of responsibility, one is already firmly in the personal realm of our conceptual scheme, where such mechanical questions cannot

even significantly be raised, let alone be relevant.

I do not intend to present a ‘solution’ to the problems of responsibility and free will here, but certainly a first step in any such solution must be

finding the crucial difference between intentional and unintentional actions. It has been a recurring theme in this book that awareness and the control of behaviour are only circumstantially related, and yet the

distinction that has so far been drawn between intentional and

unintentional actions is one of awareness: one is aware1 of the efferent commands in cases of intentional action. The awareness line is no centre of personal control, so it can hardly be that the importance of intentions has to do with a person’s control over his own motions. It could, on the other hand, be that the importance of intentions has to do with the control or influence *another* person can have over a person’s actions. The concept of intentional action hinges on the effect on people of verbal stimulation.

Verbal stimulation – talking to someone – contributes to the control of behaviour in much the same way non-verbal stimulation does. Efferent

signals have been likened to orders or commands, but a verbal order,

telling someone to do something, does not have the same function as such an efferent signal. It is a bit of information the contribution of which depends on the pre-existing neural organizations and states; the order may be obeyed or disobeyed. In extremely docile or dependent people, or in the face of over-powering authority or when one is caught off guard, verbal stimulation may in fact contribute so strongly to the determination of

behaviour that it is like pushing a button. In such a case it would be

tempting to say the verbal order causes the action, just as the efferent order causes the action, but there is a difference. The verbal order’s

content is determined by its linguistic parts, and it may or may not have the effect it is ‘designed for’, depending on conditions in the recipient.

The effects of suggestions, requests, reports of information and criticisms are similarly dependent on these conditions.

In order for verbal stimulation, as for non-verbal stimulation, to

contribute to behaviour control, the behaviour in question must be

amenable to influence and the stimulation must be relevant. Verbal

stimulation must strike at the actual controls of behaviour if it is to have its effect, and in some sorts of behaviour, verbal stimulation is simply not effective. There is little or nothing one can say to stop a person from crying, and nothing one says will stop a person from shivering. When

actions are intentional, on the other hand, the actor can report – fallibly, but normally reliably – some part of the controls, and these reports allow others to aim verbal stimulation with some assurance of accuracy and

efficacy. Suppose, for example, the lady next door is yelling at the top of her lungs and I wish to change her behaviour. Setting aside the sub-human alternative of physically muffling her, my first move is to find out *why* she is yelling. Such verbal stimulation as suggesting she install a telephone will miss the mark if the description under which her noisy behaviour is an intentional action is ‘rehearsing “Vissi d’arte” ’. Having learned the

description of her intentional action, a number of alternatives are open to me, depending on my tact and subtlety, and if I learn that an anvil has just fallen on her foot, I will abandon verbal stimulation and set about finding medical aid.

Anscombe says, ‘Roughly speaking, it establishes something as a reason

if one argues against it.’[10 T](#)his is the basis of importance of intentional actions: they are actions one can argue against. We exculpate the insane on the grounds that rationally directed verbal stimulation fails to have its proper effect: ‘It’s no use talking to him – he’s mad’, ‘He won’t listen to reason’, ‘Arguing will get you nowhere’. *To argue with an entity is to treat that entity as a person, a rational agent*. Thus personal responsibility –

and only people are responsible – is founded on the general assessment of the limits of the contributions one can make to another's behavioural

control by means of rational discourse. Anscombe argues that 'the concept of voluntary or intentional action would not exist, if the question "Why?", with answers that give reasons for acting, did not.'11 The ordinary concept of intention, which is dependent on the concepts of awareness and rational control, and on which depends the concept of responsibility, is important because for the average man the best hope he has of contributing to the control of another's behaviour lies in aiming verbal stimulation accurately at those controls that can be altered. The ascendancy of this method has until recently been supported by results, and is entrenched in our

conceptual scheme in such concepts as *rational agent*, *conscious act*, and ultimately, *person*. A shift to other methods, including non-rational verbal stimulation (hypnotism, psychoanalytic therapy, brainwashing) and

possibly chemical and electrical methods, could result in a shift in this part of our conceptual scheme, and of course this eventuality has already been dimly discerned by the users of such neologisms as

'depersonalization' and 'dehumanization' and in the prophecies of a

disappearance of the concept of responsibility under the new wave of

'causal' explanations of human behaviour in the social sciences. The

distinction between manipulation and persuasion is thus fundamental to our conceptual scheme, since on it rests ultimately the concept of a person.

This is the 'different light' in which we view motions when we view them Intentionally as actions.

10

LANGUAGE AND

UNDERSTANDING

XXIV KNOWING AND UNDERSTANDING

Viewing a person as an Intentional system is viewing a person as working with information, knowing facts, believing statements. In Chapter 4 a

number of obstacles were placed in the way of the centralist programme of associating verbal formulae (reports, statements of fact, commands, etc.) with events and states of such a system, and it might seem to be a

corollary of this that the centralist, working his way up from the sub-

personal, physical account to his merely approximate ascriptions of

content, must necessarily fail to achieve the *precision* with which we speak at the personal, purely Intentional level of people's beliefs and knowledge. The precision we find in our ordinary talk of beliefs and

knowledge is an illusion, however, for the obstacles that face the centralist have their counterparts on the purely personal, Intentional level.

We talk about what a person knows as if we could make a list of the

things he knows, or at least specify quite precisely a few of the things he knows, but such specifications as we can make are always open-ended and depend on an indefinite number of assumptions. To bring out this

systematic impossibility of precisely determining things known we must

first set aside a difficulty that infects our ordinary concept of knowledge.

The ordinary use of 'know' carries with it the claim that what is known is true. If I claim to know that *p*, and *p* turns out to be false, my claim to knowledge is disallowed. It will be said that I only believed that *p*, but did not know that *p*. Yet at the same time we suppose that what is known by a person can occupy a special psychological position, so that a person can tell the things he knows from the things he merely believes, can follow such maxims as 'Don't commit yourself until you know for sure', can tell us what he knows about a particular subject. These two notions about

knowledge, the truth condition and the ability of the person to tell

knowledge from belief, are incompatible. No one could ‘intuit’ or

‘introspect’ a difference between those things merely believed and those things believed and actually true. There can be degrees of belief; a person can order a group of statements according to how willing he would be to stake something on them – money, or his life or reputation – and he could even decide to draw a line somewhere dividing what he claims to be

knowledge from what he claims to be mere belief, but such a line must be

arbitrary. When called upon to produce one’s knowledge one can do no better than to produce what one believes to be true, and whether or not what one believes to be true *is* true does not affect its being one of those things one will produce as knowledge when asked, or will otherwise act on *as if one knew them*. If we suppose for the moment that it is safe to think in this way of *things* (perhaps facts or propositions) one will act on as if one knew them, then it is clear that the class of these things for a person need not (and probably never does) coincide with the class of things

known by the person (which, however else it is characterized, will include the truth condition). A thing (a fact or proposition or whatever) could not occupy a special psychological position (e.g., have a special functional potential in the direction of behaviour) in virtue of its truth, so knowing something cannot be purely a matter of being in a particular psychological state. When someone claims to know something and is proved wrong, it

would be absurd for him to suppose that he had misidentified the state he was in, had mistaken the marks of belief for the marks of knowledge. It is easy to confuse these two ‘classes of things’, the psychologically

characterized class of things that are *as if known* by a person with the class of things actually known. For example, characterizations of knowledge

often specify a condition of ‘adequate evidence’ or ‘justification’. What is known must pass certain tests, but which tests? If what is known is to be distinguished from what is merely believed (even believed on good

evidence or with good reason) then these tests that must be passed must be foolproof tests – but of course there are no such tests. What is *as if known*, on the other hand, must merely pass whatever tests a person sets, whether these are good or bad, wise or foolish; it is what has arrived at a certain functional position regardless of the rigours of its journey. Stipulating conditions of adequate evidence or reason is a mistake, for only 'perfect'

evidence would *ensure* knowledge; such stipulations should be construed as normative: one should test one's beliefs rigorously if one wants to

achieve knowledge.

Determining precisely what a person actually knows would involve two

tasks: first determining what a person 'knows' (what he will offer or

exhibit as knowledge), and then determining which of this is true. The

second task does not concern us here, for the issue is whether there is any way of describing or expressing or determining precisely what a person

'knows'. We are inclined to think of a person as having a store of

information and misinformation, and the question is whether we can specify the contents of this store with any precision.

Storage of information (including misinformation) does not by itself

constitute 'knowing' (I shall drop the scare-quotes from this non-ordinary term in what follows); dictionaries and encyclopaedias and libraries are stores of information, but they do not know the information stored.

Knowing requires understanding, and here we must be careful to

distinguish different sorts of understanding. One can understand each word in a sentence without understanding the sentence ('Sufficient unto the day is the evil thereof ' is one that puzzled me as a child even though I

understood each word); one can understand a sentence without

understanding what a person is saying or stating by uttering the sentence; one can understand a person's statement without understanding the person; one can understand a subject, a state of affairs, a problem.

If a person utters a sentence and we take this as an indication that the person knows (or believes) what the sentence states, we assume that the person understands the statement he has made in uttering the sentence.

What is involved in understanding a statement? It is not enough that the person be able to produce paraphrases of his sentence. A computer

programmed to translate English into Russian might be capable of

producing passable Russian paraphrases of English sentences, but this

would not suffice to show that the computer understood statements. It

might even produce English paraphrases of English sentences. Suppose we feed in 'I just murdered my uncle'. Even if it responded with 'You have recently slain the brother of one of your parents' it would not be said to have understood the statement. If, however, the computer immediately

made a discreet telephone call to police headquarters, one would be

tempted to say it had understood the statement, but only if it also had the capacity to do other quite different things with different input. If it is merely the local ADIAC computer (Apparatus of Dubious Intelligence for

Acknowledging Confessions) no one will grant it understanding. Only a

being that is non-verbally active in the world could meet our requirements for understanding. A computer whose only input and output was verbal

would always be blind to the meaning of what was written. It might

'grasp' all the verbal connections, but it would lack 'acquaintance' with the things the words are about. Suppose we fed a computer a description of the Taj Mahal. It might paraphrase this or even respond with an output like

‘The Taj Mahal must be very beautiful’, but one wants the computer also to produce outputs like ‘Take me there; I want to see for myself ’, and such outputs would be a hoax if the computer did not have some perceptual

apparatus and many other sophisticated capacities.

The tests for understanding in each particular case involve a ‘family’ of behavioural capacities, some of which must be demonstrated. If Jones says

‘Smith is here’ no one will allow that Jones understands this, and hence no one will allow that he knows or believes this unless Jones can also say and do a variety of other things ‘with his knowledge’. He must be able to

assert, for instance, ‘Smith is not in Siam’, ‘Smith is that friend of

Black’s’, or ‘By “here” I mean “in town”, not “in this room” ’. If Jones knows Smith is here he must be able to point him out or at least direct the search party. If no such corroborating behaviour is in the offing, Jones may be in this instance no more than a parrot or tape recorder. The particular tests that must be passed in any one case are not entirely determined by the information in question (the candidate for what is known). Rather,

what tests must be passed depend largely on what *else* the person knows and understands, and whereas a great deal of the corroborating behaviour can be verbal – explaining, asserting related statements, paraphrasing, and expanding on the subject – if there are available non-verbal tests and they are failed, the verbal testimony will be shaken. Much information, of

course, is so intimately verbal in being about verbal states of affairs that no strictly non-verbal behaviour could tend to corroborate the claim to knowledge: e.g., the information that yesterday was (called) Friday, this place is (called) New York.

What are the conditions that would suffice to show that a child

understood his own statement: ‘Daddy is a doctor’? Must the child be able to produce paraphrases, or expand on the subject by saying his father cures sick people? Or is it enough if the child knows that Daddy’s being a doctor

precludes his being a butcher, a baker, a candlestick maker? Does the child know what a doctor is if he lacks the concept of a fake doctor, a quack, an unlicensed practitioner? Surely the child's understanding of what it is to be a doctor (as well as what it is to be a father, etc.) will grow through the years, and hence his understanding of the sentence 'Daddy is a doctor' will grow. Can we specify what the child knows when he tells us his Daddy is a doctor? It may seem simple: what the child knows is that his Daddy is a doctor – that is, the object or content of his knowledge in this case is the

proposition, 'that Daddy is a doctor'. But does the child *really* know this?

One is inclined to say that he only 'sort of ' knows this, or 'half ' knows this. If the proposition is to be the thing known, we have to allow for quasi-knowledge of propositions. Yet one might argue that when the child only half knows the proposition there is still something – something

somehow 'less' – that he fully or really knows.

If understanding admits of degrees, then so must knowledge, since

understanding is a condition of knowledge, and this bodes ill for *things* known, for facts, or propositions or whatever. A child with the rudiments of arithmetic under his belt knows what the number four is, or knows a

little bit about the number four. The effect of this small knowledge is that the child can reel off reports of his knowledge *ad nauseam*: 'four is half of eight …, four is 1/250,000 of a million'. None of these reports, and no finite collection of them, exhausts his knowledge of the number four; and since he is not an advanced student of mathematics we cannot expect him to offer, assent to or even understand statements about the real-number system, the infinite multiples of four – in short the sort of statements that might be held to generalize and exhaust his knowledge. What fact or facts can we say the child knows? Does he know an infinite number of facts or just one or two rather general facts? If the latter, he knows facts the

'expression' of which in his native tongue he probably does not even

understand. And what of the child whose arithmetical knowledge is shaky, who knows that two times four is eight, three times four is twelve, but who is unsure about whether four times two is eight, four times three is twelve, or who denies the latter equations? If he does not have the symmetry of

'times' down pat, can we really say he knows that two times four is eight, and if this is not what he knows, what does he know?

Ryle has drawn his well-known distinction between knowing how and

knowing that. When one claims to know how to swim, one supports the

claim by swimming, not by an essay on swimming methods; knowing how

is a matter of a talent or knack or ability, not a matter of having some propositions or facts in one's head. The distinction is certainly

illuminating at some levels of discussion, but does it break down when it comes time to determine just *what* is known when one knows *that* something is the case? The case of the child's arithmetical knowledge is revealing because it seems to straddle the line between know-how and

know-that. The child's knowledge is very much like a knack or trick he has

learned, and yet it is more; what the child does in demonstrating his knowledge is more than merely a rote parroting, the utterance of a string of phonemes. Yet what the child demonstrates is apparently more like a

knack than knowing that yesterday was Friday. Where does one draw the

line? Said of an adult that for any (true) *p* either he knows that *p* or he does not is initially plausible until we examine the penumbral cases. 1 Do I know that table salt is sodium chloride? Yes, of course. But do I? What does this mean to me, or what can I do with this information beyond just reporting it? Educators are prone to distinguish the mere learning of facts from understanding or learning with comprehension, but do we want to say a person actually knows a fact if all he can do is utter some sentence in a few limited contexts (in response to an examination question, for

example)? Here we seem to have just the ‘opposite’ of knowing that – a clear case of rote know-how. Surely what the trained chemist knows when he knows that table salt is sodium chloride is more than what I know.

Imagine stationing a man who does not understand German on a street corner and training him to respond to ‘Wo ist der Bahnhof?’ with ‘Der Bahnhof ist links um die Ecke’. Is not my knowledge that salt is sodium chloride merely on a sliding scale from his ‘knowledge’ that der Bahnhof ist links um die Ecke? Understanding the statement that salt is sodium chloride involves more, of course, than just understanding the words, or ‘knowing their uses’; understanding the statement involves knowing about sodium and chlorine, but also about potassium and oxygen and valences and so forth. We cannot draw a limit so that understanding a statement involves understanding just so much.

If one can talk of a fact known, it must vary from speaker to speaker for any given sentence which might be held to express a fact. Alternatively, if one can settle on some way of anchoring facts to sentences, then these will not serve well as *things known*. An encyclopaedia – a very small one – might be held to store just one fact if it consisted of one non-compound printed sentence, but a person could not be held to know just one fact or understand just one statement. The knowledge of one fact could not exist by itself because the fact could not be used, and hence could not be understood. However facts may be anchored for encyclopaedias, the metaphor of the walking encyclopaedia is not to be trusted unless what we mean is just that the ‘walking encyclopaedia’ really *knows nothing*, is quite literally no more than a sort of recorder. What a person can use his

stored information for depends on what other stored information he has, what else he knows. The things we do with our knowledge are quite

discrete, but our knowledge itself does not divide into neat, independent parts, and hence cannot be ‘listed’.

XXV LANGUAGE AND INFORMATION

Philosophers of language in the past have attempted to pin down the

information ‘contained’ in particular sentences by appealing to particular placements of things in the universe (the cat on the mat), particular time-slices of spatio-temporal reality, particular concatenations of qualities present to the senses, but these attempts to tie information to states of affairs of one sort or another fail because they do not take into

consideration the intermediaries between sentences and states of affairs, namely the sentence utterers and hearers, the *makers* of verbal messages.

A message picks out some feature of a state of affairs that is functionally important to some ‘receiving system’; something is a message or a signal only when it goes on to effect functions in some self-contained Intentional system. The freezing of a pond is not in itself a signal to the effect that the temperature of the water is below the freezing point.

In information theory it is often important to gauge the reliability of an information transmission channel or system, and for this a method has

been developed for measuring *amounts* of information. Significantly, the amount of information in any signal is not directly a function of stimulus conditions or causes of the signal, or of any internal syntactic structure of the signal (which is usually, in any case, treated holistically as an ‘off’ or

‘on’ in a binary system). Rather the amount of information is determined by the degree of uncertainty diminished in the receiver. The receiver is given the task of singling out some individual or individuals from a

limited ensemble or class of possibilities, e.g., finding out what day of the week it is. The signals received serve to exclude possibilities (e.g., the signal ‘It is not a weekday’), thus reducing the ensemble, or one signal can single out the individual; solving the problem in one step.

How much information is in the statement ‘This is Friday’? We

now know that we must first determine the context. Suppose our

ensemble was ‘The days falling between Thursday and Saturday’.

Such an ensemble has one member, so that

I = log2 1 bits [the unit of information]

= 0 bits

The statement, then, contains no information.

In another context the result could be different. Suppose we know that since we are working it is neither Saturday nor Sunday.

In this case, our ensemble has five equiprobable members, and

I = log2 5 bits

= 2.32 bits

Finally, let us suppose a man awakens from a coma. He has no

idea how long he has been unconscious, and asks ‘What day is it?’

The seven possible outcomes are equiprobable, and

I = log2 7 bits

= 2.81 bits. 2

A signal or message, then, like ‘This is Friday’, informs only relative to its function in ordering an ensemble, and *the ensemble is determined by the receiver*. To the man who knows that yesterday was Thursday and

tomorrow is Saturday, ‘This is Friday’ is no news. This way of

determining the amount of information works only for ensembles with a

known number of equiprobable members. Thus it is of no use in

determining the information content of ‘Your uncle just died’ or most of the sentence tokens occurring in everyday life. In the case of ‘Your uncle just died’ the ensemble might be held to consist of two members, uncle

dead or alive, and if a person were waiting for news on the state of his uncle, then to say that the sentence carries log22 or 1 bit of information would make some meagre sense. But in human beings, as opposed to

devices with one limited job to do, the receipt of information allows a great many different ensembles to be partially ordered, depending on the knowledge already held by the receiver. Thus one of our intuitions about information in people is provided with a quasi-mathematical model: the

information received by people when they are spoken to depends on what

they already know and is not amenable to precise quantification.

An intuitive account of what happens when I tell someone something is

that I try to transmit or impart or share with my listener something

somehow held in me (as known or believed); I try to produce in another

person something (knowledge of something) that I have. What is perfectly clear is that this *something* I am trying to transmit is information, and the *something* I am trying to produce in the other person is the storage of information, of information having the same or similar content to

information stored in me. However, as soon as one supposes that sentences

uttered are straightforward vehicles of particular and determinate morsels of information, that out of the building blocks of language we can

construct vehicles of just so much information on a particular topic,

paradoxes arise. If I do not know that Tully is Cicero and announce

'Cicero denounced Catiline', my listener, if he knows Tully is Cicero, will in effect come to know more than I was endeavouring to tell him. He will receive and store not only the information that Cicero denounced Catiline, but also the information that Tully denounced Catiline. Should one say

that he received more information than I sent? Such dividends of

information do not always hinge on synonymy or identity of reference of terms as in the Tully-Cicero case. On hearing 'Your uncle just died' Jones may be informed that he will soon be a rich man, that a certain Mrs. Smith is now a widow, etc., and however information is to be construed these

dividends are not the *same* information the speaker intended to impart.

If, in telling someone what I know, the effect that signals success is the production of information storage in him similar or the same in content to information stored in me, the chances of success seem remote. If, as I

argued earlier, content is in part a function of capacity to direct further efferent activity, it would be very rare for the listener to acquire the same content as that stored in me, since there will always be differences in these capacities except in the unlikely case where the listener has an information store that already duplicates mine in every relevant respect save just what I am communicating to him. This is brought out in a definition of meaning provided by MacKay. He defines 'the meaning of an utterance (intended,

standard, received) as its selective function (intended, standard, actual) on the range of possible states of the appropriate system'. 3 As MacKay points

out, this makes meaning relative, not an absolute property. If intended meaning is to be at all an approximation of standard meaning or actual

meaning, there must be some similarity in selective function of the

utterance from person to person, and this will only be the case if there is considerable similarity from person to person of the information storage systems. This requirement is easily overlooked, but it is evident in some of our everyday observations about our successes and failures at

communication. Both speaker and hearer must share relevant knowledge

for communication to occur. There is more to this than the fact that if I attempt to communicate in English with a person who speaks no English, I will not succeed in producing in him information similar to mine at all.

Even if the hearer is English, he must also have much the same background of information on the subject of discussion as I have. The

sentence ‘I’ve found a solution to the problem of other minds’, which

contains no words that the average adult English speaker would not know, is still unlikely to be informative to a person who does not share with the speaker a background of knowledge of this traditional philosophical

problem, the speaker’s activities, and what might be considered to be a solution to the problem. As MacKay says, ‘topic understanding multiplies the informational impact of a proposition’. 4 The similarity of background information stored need not be complete, and where it is partial,

communication is partial, as we observe when we say ‘I understand you,

but what you say doesn’t tell me very much’.

The fact that the information-bearing capacity of language is thus

finally dependent on the effects of language on a *person*, on what I have called an Intentional system, is the same fact on the personal level as the

fact on the sub-personal level that the centralist is unable to ascribe precise contents to the events and states of such a system.

XXVI CONCLUSIONS

The problem of mind is not to be divorced from the problem of a person.

Looking at the ‘phenomena of mind’ can only be looking at what a *person* does, feels, thinks, experiences; minds cannot be examined as separable entities without leading inevitably to Cartesian spirits, and an examination of *bodies* and their workings will never bring us to the subject matter of mind at all. The first step in finding solutions to the problems of mind is to set aside ontological predilections and consider instead the relation between the mode of discourse in which we speak of persons and the mode of discourse in which we speak of bodies and other physical objects. This studious avoidance of ontological commitments allows us to relax the

requirements of a *rapprochement* between the language of mind and the language of science, and, as we have seen, none of the freedom provided us by this stance is gratuitous. Thoughts, for example, are not only not to be identified with physical processes in the brain, but also not to be

identified with logical or functional states or events in an Intentional system (physically realized in the nervous system of a body). The story we tell when we tell the ordinary story of a person’s mental activities cannot be mapped with precision on to the extensional story of events in the

person’s body, nor has the ordinary story any real precision of its own. It has no precision, for when we say a person knows or believes this or that, for example, we ascribe to him no determinable, circumscribed, invariant, generalizable states, capacities or dispositions. The personal story,

moreover, has a relatively vulnerable and impermanent place in our

conceptual scheme, and could in principle be rendered ‘obsolete’ if some day we ceased to *treat* anything (any mobile body or system or device) as an Intentional system – by reasoning with it, communicating with it, etc.

That day is not to be expected – and certainly not hoped for – in spite of the inroads that are now being made in ‘impersonal’ ways of controlling people.

The feature that is central (if not quite universal) in the personal mode of discourse is Intentionality, and it is this feature that persistently tempts the theory-builder into positing mananalogues as elements in his analysis, thus obviating the analysis entirely. In his purest form the little man in the brain takes on the guise of brain-writing reader, an intelligent,

communicating system capable of understanding messages. Positing the brain-writing reader is almost irresistible, for if we cannot understand central states and events of the nervous system as bearing content, as

being messages of some sort, it is not clear how we can understand them at all. The temptation must be resisted, however, by recognizing the

disanalogies between verbal communication and non-verbal intra-cerebral communication and indeed the primacy of non-verbal communication.

Other roles played by the little man in the brain are merely specialized roles projected inwards from the details of our initial analysanda, the variety of affairs of a person. The solitary audience in the theatre of consciousness, the internal decision-maker and source of volitions or

directives, the reasoner, if taken as *parts* of a person, serve only to postpone analysis. The banishment of these concepts from our analysis

forces the banishment as well of a variety of other self-defeating props, such as the brain-writing to be read, the mental images to be seen, the volitions to be ordered, and the facts to be known. These props are self-defeating because they could only serve the functions for which they were designed in conjunction with interior person-analogues, and hence as

elements in an analysis they reproduce the problems like images in a hall of mirrors.

NOTES

1

THE ONTOLOGICAL PROBLEM OF MIND

1 In the immense literature of identity theory, several items stand out as germinal: U. T. Place,

‘Is Consciousness a Brain Process?’, *British Journal of Psychology*, XLVII, 1956, pp. 44–

50. H. Feigl, ‘The “Mental” and the “Physical” ’, *Minnesota Studies in the Philosophy of Science*, Vol. 3, eds. H. Feigl *et al.* , Minneapolis, 1958, pp. 370–457. J. J. C. Smart,

‘Sensations and Brain Processes’, *Philosophical Review*, LXVIII, 1959, pp. 141–56. In the spate of papers following these the greatest advance of outlook is to be found in T. Nagel,

‘Physicalism’, *Philosophical Review*, LXXIV, 1965, pp. 339–56. Place’s and Smart’s papers are reprinted with revisions in V. C. Chappell, ed., *The Philosophy of Mind*, Englewood Cliffs, 1962.

2 G. Ryle, *The Concept of Mind*, London, 1949.

3 I must postpone briefly an important and difficult question: are sentences containing these contextual disharmonies such as ‘I can sit on an opportunity’ syntactically ill-formed (and hence neither true nor false) or are they false by meaning (and hence have true negations)?

For an excellent discussion of this and other questions raised in this section, see F.

Sommers, ‘Types and Ontology’, *Philosophical Review*, LXXII, 1963, pp. 327–63, reprinted in P. F. Strawson, ed., *Philosophical Logic*, Oxford, 1967.

4 W. V. O. Quine, *Word and Object*, Cambridge, Mass., 1960, p. 244.

5 *Ibid.* , p. 244.

6 Theorists of space and measurement have not always been alive to this category mistake.

Descartes, for one, advanced the argument against the existence of vacuums that if there was no matter between A and B there could be no distance between them (*Principles of Philosophy*, Part II, secs. 16–18).

7 The reader is invited to form other natural sentences affirming the existence of voices. He will find that we very rarely assert flat out the existence of a voice or voices. In this respect a voice has a much slimmer claim to reality than, say, a twinkle in the eye.

8 See note 3 above.

9 This position differs substantially from Ryle's. He supposes that it will suffice to talk of different types or categories of existence (*op. cit.* , pp. 22ff.). With regard to the view proposed above, it cannot be denied that there is also something counterintuitive in holding these odd sentences to be ill-formed, not logically false. For if there is something queer about 'I can sit on an opportunity', it seems that a very natural, ordinary, oft-heard way of alluding to this queerness would be to say: 'but you *can't* sit on an opportunity; it isn't that sort of thing.' This must also be ill-formed, however, in spite of intuitions. Such locutions can be kept as short cuts, however, for the more proper: ' "Opportunity" is not accepted into the context "can sit on …" ', and I do this several times in this chapter.

10 My use of 'referential' is only close to that of Quine in *Word and Object*, but I chose the term for these affinities. Another term, 'syncategorematic', is close to my 'non-referential', but was rejected since in its established use it generally describes adjectives, not nouns, and

stresses class determination over existence; thus 'expectant' is syncategorematic in

'expectant mother' since the class of expectant mothers is not the subclass of mothers who are expectant (see Quine, *op cit.* , p. 103). The distinction I wish to mark is not that between

‘horse’ and ‘centaur’ (there are horses, but no centaurs; in that sense ‘centaur’ does not refer), but that between ‘horse’ and ‘when’; ‘when’ is non-referential in my sense, which does not mean that *whens* are mythical or extinct; Quine would say ‘when’ was *not a term*.

11 Karel Lambert has tried to persuade me that an adequate logical language for maintaining the ontological neutrality I require here would be a two-quantifier logic such as those developed by Leonard, Van Fraassen, and himself (H. S. Leonard, ‘Essences, Attributes and Predicates’, *Proceedings and Addresses of the American Philosophical Association*, Vol.

XXXVII, 1964, pp. 25–51. B. Van Fraassen, ‘Meaning Relations among Predicates’, *Nous*, I, 1967, pp. 161–79. R. K. Meyer and K. Lambert, ‘Universally Free Logic and Standard Quantification Theory’, *Journal of Symbolic Logic*, XXXIII, No. 1, 1968, pp. 8–26. B. Van Fraassen and K. Lambert, ‘Quantifiers, Meaning Relations and Modality’ in K. Lambert, ed., *Philosophical Development in Non-classical Logic; Modality, Existence and related areas*). These languages were developed to deal with rather different problems, in particular the ‘possible objects’ so handy to modal logicians, and adapting my position to them would require allying voices, thoughts and minds to centaurs and gryphons (if not to round squares), and this would clearly be a distortion of the view I present. Whether in the end it would be a logically or philosophically undesirable distortion is not yet clear to me.

12 In other words we start by treating the sentences in the fashion of the p and q of the propositional calculus; what then confronts us is whether their parts can also be brought under the quantifiers of the predicate calculus (viewed, in Quine’s fashion, as ontologically committing).

13 H. Putnam, ‘Psychological Predicates’ in W. H. Capitan and D. D. Merrill, eds., *Art, Mind and Religion*, Pittsburgh, 1967, pp. 37–48. It might be argued that I have been unfair to the identity theorists, and that Smart and others have in fact made use of the ontological points I have raised. Smart, after all, in response to the objection that an after-image was not a brain process, replies ‘I am not arguing that the after-image is a brain process, but that the experience of having an after-image is a brain process’ (‘Sensations and Brain Processes’, revised version in Chappell, *op. cit.* , p.

168). One might interpret this as holding that ‘after-image’ was not the referential atom, but rather ‘having-an-after-image’, a fused idiom.

Nagel (*op. cit.* , p. 341) goes even farther: ‘Instead of identifying thoughts, sensations, after-images, and so forth with brain processes, I propose to identify a person’s having the sensation with his body’s being in a physical state or undergoing a physical process.’

Nagel’s position is very close to mine in that he has simply taken a mental language sentence whole, turned it into the appropriate gerund nominalization, and put it next to an identity sign flanking a similarly altered physical entity sentence. But Putnam’s objection holds as well against Nagel’s more circumspect identities; the correlation Nagel supposes must hold is still too strong to be plausible. Moreover, hasn’t one lost the point of *identity* theory once one begins treating whole sentences as names in effect of situations or states of affairs which are then proclaimed identical with other situations or states of affairs?

2

INTENTIONALITY

1 I shall capitalize Brentano’s term and its derivatives to distinguish them from the somewhat related and etymologically similar family of more common terms, ‘intend’, ‘intentions’,

‘intentionally’, which will be examined in detail in Chapter 9. It seems clearest to carry this practice into quotations as well.

2 F. Brentano, *Psychologie vom Empirischen Standpunkt*, Leipzig, 1874, Vol. I, Book II, Chap. i, ‘The Distinction between Mental and Physical Phenomena’ – a selection in R.

Chisholm, *Realism and the Background of Phenomenology*, Glencoe, 1960, trans. D. B.

Terrell.

3 Brentano called these objects ‘fictions’, indicating his own refusal to take seriously a metaphysical class of ‘inexisters’. For an excellent discussion of Intentional objects, see G.

E. M. Anscombe, ‘The Intentionality of Sensation: A Grammatical Feature’, Sec. 1, in R. J.

Butler, ed., *Analytic Philosophy* (Second Series), Oxford, 1965, pp. 158–68.

4 R. Chisholm, *Perceiving: a philosophical study*, Ithaca, 1957, p. 170.

5 *Ibid.* , pp. 170–1.

6 Chisholm himself has attempted in a number of papers to reformulate his criteria to meet objections. (See especially his ‘On some psychological concepts and the “logic” of Intentionality’, in H. Castañeda, ed., *Intentionality*, *Minds and Perception*, Detroit, 1967, pp. 11–57.) Since our aims diverge, however, there is no point here in recounting his various modifications.

7 C. Taylor, in *The Explanation of Behaviour*, London, 1964, Part II, provides an excellent survey of behaviourists’ so far fruitless efforts to produce ‘operational’ definitions of such terms as ‘desire’.

8 (6) meets (3) however, since if Smith, unbeknownst to John, is his uncle, the substitution will make the sentence false.

9 Cf. S. Körner, *Experience and Theory: an essay in the philosophy of science*, London, 1966, p. 200.

10 P. T. Geach, ‘Intentional Identity’, *Journal of Phil.* LXIV, 20, 1967, p. 629.

11 Quine, *op. cit.* These and kindred examples are discussed in his Chs. 4–6.

12 The hope of modal logic is that it will be possible to adjust the entities referred to in modal statements, or restrict the ways in which entities may be described, so that substitution –

where it is permitted – still preserves truth. For example, one might claim that the only sense of ‘9’ that renders (10) true renders the identity ‘9 = the number of planets’ false; or in the same vein, that the sense of ‘the number of planets’ needed to make the identity true makes (11) true as well (a distinction being seen between (11) and ‘there are necessarily more than 7 planets’.

13 The situation is not that simple, however. In the view of some modal logicians, any success with the alethic modalities will bring in its wake parallel successes in the other modalities, *including* the psychological or Intentional modalities. In this event the logic of Intentionality would be merely part of a unified modal logic, and instead of modal logic providing a solution to our counter-example, it would go on to provide something very like a refutation of the Intentionalist thesis of non-reducibility. This eventuality seems remote to me, as should become clear in what follows.

14 See Quine, *op. cit.* , Ch. 6.

15 Chisholm himself accepts the conclusion that (1)–(3) do not serve to distinguish the psychological from other modalities, and attempts to make the distinction in another way (Castañeda, *op. cit.* , p. 11).

16 Cf. *ibid.* , p. 33.

17 For an example of this sort of difficulty, see my ‘Geach on Intentional Identity’, *Journal of Philosophy*, LXV, II, 1968, pp. 335–41.

18 R. Carnap, in *Meaning and Necessity*, Chicago, 1947, uses ‘intensional’ in a narrower sense to refer to a subclass of non-extensional sentences.

19 Cf. Taylor, *op. cit.* , p. 200.

20 Quine, *op. cit.* , p. 220.

21 As a response to the request to utter a three-word sentence in English, saying ‘it is raining’

would be a case where (16) was true and (17) false.

22 Quine, *op. cit.* , pp. 220–4; Chisholm, *Perceiving*, Ch. 11.

23 Cf. Körner's distinction between dogmatic and methodological behaviourists (*op. cit.* , Ch.

XIII). It is the difficulty with Intentional language that turns methodological behaviourists into dogmatic behaviourists, but Körner claims this is not the only refuge.

24 Quine, *op. cit.* , p. 221.

25 K. S. Lashley, *Brain Mechanisms and Intelligence*, Chicago, 1929.

26 There is an incisive and systematic survey of the frustrations of the behaviourists in Part II of Taylor's *Explanation of Behaviour*. Although, as he points out, his critique is open-ended, the troubles so far encountered exhibit 'fairly reliable signs' of futility. In particular there is the accumulation of inelegant *ad hoc* props and provisos that make stimulus-response theories so stupefyingly complex (p. 272). D. Shwayder's *Stratification of Behaviour* analyses similar obstacles to stimulus-response theory, but Shwayder does not hold it to be *in principle* impossible for the behaviourists to produce a theory complex enough to deal with the problems he raises.

27 For example, an animal in a problem-solving experiment will be 'trained to criterion', which means: given enough training trials to bring his performance up to an arbitrary standard of success, say 9 out of 10. But 9 out of 10 what? The 'strictly behavioural'

criterion of learning is in actuality hedged with a *ceteris paribus* clause; what it means is: 9

times out of 10 the animal *finds* the goal box it was *looking for*, it *achieves* its goal.

28 There could be, of course, 'autonomous sciences of Intention' without all the trappings of the Phenomenological movement, but with many centrally shared features.

29 Arguments designed to show roughly this are found in Taylor, *op. cit.* , A. I. Melden, *Free Action*, London, 1961 and Anscombe, *Intention*, Oxford, 1957.

30 Cf. Taylor, *op. cit.* , p. 44.

31 Taylor says: ‘The notion of adaption is of course implicit in the ordinary language teleological form of explanation where action is frequently explained in terms of its propitiousness for certain purposes, i.e. by its “adaptiveness” in respect of these ends. The aim of S-R theories on the other hand is to explain behaviour without using a notion of this kind. Thus for Hull it is one of the tasks of a molar science of behaviour to explain why behaviour is adaptive, why “it is successful in the sense of reducing needs and facilitating survival”, a task separate from though closely related to that of explaining why the behaviour of different organisms is as it is. Adaptiveness is thus an explicandum for S-R

theory. It is not a principle to be used in the explicans’ (*op. cit.* , p. 115). Cf. I. Sheffler, *The Anatomy of Inquiry*, New York, 1963, p. 92.

32 C. S. Pittendrigh, ‘Adaptation, Natural Selection and Behavior’ in A. Roe and G. G.

Simpson, eds., *Behavior and Evolution*, New Haven, 1958, p. 395.

33 Cf. Taylor, *op. cit.* , pp. 107–8. Taylor points out (p. 271) that Freud was a centralist, hoping (rather faintly) to find a physiological basis for such Intentional phenomena as (subconsciously) intending to do something, fears, hates and repressed memories.

3

EVOLUTION IN THE BRAIN

1 D. M. MacKay, ‘Towards an Information-Flow Model of Human Behaviour’, *British Journal of Psychology*, XLVII, 1956, pp. 30–43. For another description of storage see Taylor, *op.*

cit. , p. 108.

2 J. Y. Lettvin, *et al.* , ‘What the Frog’s Eye Tells the Frog’s Brain’, *Proceedings of the Institute of Radio Engineers*, 1959, pp. 1940–51.

3 The example is D. Wooldridge’s in *The Machinery of the Brain*, New York, 1963.

4 B. F. Skinner, *The Behavior of Organisms*, New York, 1938.

5 M. Arbib, *Brains, Machines, and Mathematics*, New York, 1964, p. 56.

6 Deprivation experiments are designed so to limit an animal’s sensory experience from birth that if the animal performs some perspicuous act, the ability to do this can be explained only as an inherited, instinctual capacity, rather than a learned one.

7 See Taylor, *op. cit.* , for a critique of these concepts, esp. pp. 170ff.

8 This formula gives a *teleological* characterization of the conditions of existence of some neural structures. The question of whether teleological explanations are eliminable in favour of non-teleological explanations is the question of whether such a formula is subject to further explanation or must just stand as it is, a brute and unaccountable fact. My claim that the formula *is* explicable in terms of the operations of natural selection – a process that can be given a non-teleological description – is thus a claim in favour of elimination of the teleological. Cf. Taylor’s discussion of teleological explanation, *op. cit.* , Part One. As Taylor also points out, the question of the elimination of the teleological is intimately bound up with the question of the reducibility of the Intentional.

9 A. Newell and H. A. Simon, ‘GPS, a Program that Simulates Human Thought’ in H. Billing, ed., *Lernende Automaten*, Munich, 1961, reprinted in E. A. Feigenbaum and J. Feldman, *Computers and Thought*, New York, 1963, pp. 279–92.

1 See, for example, D. H. Hubel and T. N. Wiesel, ‘Receptive Fields, Binocular Interaction and Functional Architecture in the Cat’s Visual Cortex’, *Journal of Physiology*, 1962, pp.

106–54; Lettvin *et al., op. cit.* and ‘Two Remarks on the Visual System of the Frog’ in W.

A. Rosenblith, ed., *Sensory Communication*, New York, 1961, pp. 757–76; W. R. A. Muntz,

‘Vision in Frogs’, *Scientific American*, 210, 1964, pp. 110–19, and D. H. Hubel, ‘The Visual Cortex of the Brain’, *Scientific American*, 209, 1963, pp. 54–74.

2 The analogy can be reversed in the case of information theory properly so called, of which parts may be considered relatively uninterpreted but amenable to a variety of different

‘meanings’ or applications or physical realizations. In this case, it should be noted, the information being considered (e.g., in terms of ‘bits’) is not *intelligently used* information.

That is, one is concerned with the reliable transmission of impulses, dots, ons and offs, letters of the alphabet, codes, but not with the understanding of messages.

3 Frank McGuinness has pointed out to me that a ‘neural negator’, for example, is a particularly unlikely bit of machinery. We cannot project grammatical transformations into the brain and hope to find transformers there. What similarities should we suppose to exist

between the stimulus sources for ‘there is danger to the left’ and ‘there is no danger to the left’? Clearly, although we can see the different sorts of *effects* these signals should have, they need share no structural similarities, nor must they both be ‘derived’ from similar stimulus conditions.

4 J. Zeman, ‘Information and the Brain’ in N. Wiener and J. P. Schadé, eds., *Nerve, Brain and Memory Models*, New York, 1963, p. 71.

5 The ‘brain-writing’ view is plagued by a host of disanalogies in any case. As D. M. MacKay *et al.* , point out, ‘The loose coupling between language and the world which distinguishes statements from symptoms is notably absent in something that has often been called a language – the representation of information in afferent nerve fibres. This is systematic but purely symptomatic, and the relations between “sender”, “user” and “referent” are considerably different. The retina cannot exercise an opinion whether to tell the brain what energy has reached it and where …’, ‘Computers and Comprehension’, *RAND Memo* RM4065PR, Apr. 1964, p. 11. (This article contains many valuable observations on the relation between understanding sentences and the analogue of understanding in computers.) See also MacKay, ‘Linguistic and non-Linguistic “Understanding” of Linguistic Tokens’, *RAND Memo* RM3892, Mar. 1964.

6 Cf. L. Wittgenstein: ‘ “And yet you again and again reach the conclusion that the sensation itself is a nothing.” Not at all. It is not a *something*, but not a *nothing* either! The conclusion was only that a nothing would serve just as well as a something about which nothing could be said.’ *Philosophical Investigations*, trans. G. E. M. Anscombe, Oxford, 1953, i. 304.

5

INTROSPECTIVE CERTAINTY

1 Wittgenstein, *op. cit.* , i. 244. See also i. 367, i. 370.

2 G. Ryle, *op. cit.* , p. 102. Ryle has since described to me a view of ‘degrees’, with ‘Ouch!’ at the avowal end of the spectrum and ‘the pain is in the third tooth, upper left’ at the other, reportorial end. This is plausible, but of course demands either an explanation of how the true reports at one end are infallible or the implausible view that only the avowals are immune to error.

3 G. E. M. Anscombe, *Intention*, p. 14.

4 H. Putnam, ‘Minds and Machines’, *Dimensions of Mind*, ed. S. Hook, New York, 1961, pp.

148–79.

5 *Ibid.* , p. 154.

6 *Ibid.* , pp. 155–60.

7 See, e.g., N. Chomsky, *Syntactic Structures*, ’s-Gravenhage, 1957, Chomsky, ‘A Review of B. F. Skinner’s *Verbal Behavior*’, reprinted in *The Structure of Language*, eds. J. Fodor and J. Katz, Engelwood Cliffs, 1964, Sec. XI, pp. 574–8; Chomsky, ‘Three Models for the Description of Language’, *I.R.E. Transactions on Information Theory*, Vol. IT-2 (Sept.

1956), pp. 113–24; also K. S. Lashley, ‘The Problem of Serial Order in Behavior’, *Hixon Symposium on Cerebral Mechanisms in Behavior*, ed. L. A. Jeffress, New York, 1951, pp.

112–36; and E. Lenneberg, ‘The Acquisition of Language’ in Fodor and Katz, *op. cit.* , pp.

579–603.

8 W. Penfield, *The Excitable Cortex in Conscious Man*, Liverpool, 1958.

9 In the past, some psychologists have wanted to hold that this phenomenon involved incipient but not quite detectable muscle movement, but although this can no doubt occur it

is not longer seen as necessary.

10 I am indebted to Dennis Stampe for raising these questions about speech acts and intentions.

11 In the light of this view, Wittgenstein’s view that ‘the verbal expression of pain replaces crying and does not describe it’ may seem on reflection to be correct, except for the suggestion it carries – due to its failure to distinguish the personal from sub-personal level –

that a verbal expression of pain (viewed from the personal level) is not intentionally a *bona fide report*, but rather an outcry of sorts.

6

AWARENESS AND CONSCIOUSNESS

1 I mean ‘speech centre’ in the functional or logical sense developed in Chapter 5, of course, and not in any standard anatomical sense. It is my impression that this definition of awareness is congenial with G. Bergmann’s remarks on awareness in *Meaning and Existence*, Madison, Wisc., 1960. If I understand him right he too is concerned with what one might call the functional location of the contents of awareness.

2 These very rough definitions leave many questions unanswered, e.g., what does ‘effective’

mean in (2), and how big a role must an event play in directing behaviour to count? Since I do not intend to build on these definitions but merely use them to illustrate a distinction, there is no point in elaborating answers to these interesting and important questions here.

As it is these unwieldy definitions can serve as reference points for the uses of the terms in what follows, but for the sake of less cumbersome expression I shall revert to the ‘aware of’

form where it is convenient, speaking of a creature being aware1 of something or aware2 of something, letting it be understood that in these cases the creature is aware1 or aware2 that *p*, where *p* is a statement informing us about the ‘object’ of awareness.

3 K. M. Sayre, ‘Human and Mechanical Recognition’ in K. M. Sayre and F. J. Crosson, eds., *The Modeling of Mind*, Notre Dame, 1963, pp. 157–70.

4 Recognition without awareness1 would be the natural way to describe what occurs when we scan a list of words for, say, a colour word, discarding all non-colour words without noticing or being aware of what they are. Clearly this must be recognition *with* awareness2

and *without* awareness1. See, e.g., U. Neisser, ‘Visual Search’, *Scientific American*, 210

(June) 1964. On *subconscious testing* and awareness, see also my ‘Machine Traces and Protocol Statements’, *Behavioral Science*, Mar. 1968, pp. 155–62.

5 M. Scriven, ‘The Mechanical Concept of Mind’, *Mind*, LXII, 1953, reprinted in *Minds and Machines*, ed. A. R. Anderson, Englewood Cliffs, 1964, p. 33.

6 K. S. Lashley, ‘Cerebral Organization and Behavior’, *The Brain and Human Behavior* (*Proc. of the Association for Research in Nervous and Mental Disease*, Vol. 36, eds. H. C.

Solomon, S. Cobb and W. Penfield, Baltimore, 1958, p. 4).

7 See, e.g., Taylor, *op. cit.* , pp. 58–67, and D. Hamlyn’s review of Taylor in *Mind*, LXXVI, 1967, p. 130.

8 Cf. Anscombe, *Intention*, p. 86, where this conflict is clearly presented but not clearly resolved.

7

MENTAL IMAGERY

1 Optimists who doubt that mental images are still taken seriously in philosophy and even in science are invited to peruse two recent anthologies, R. J. Hirst, ed., *Perception and the External World*, New York, 1965, and J. R. Smythies, ed., *Brain and Mind, Modern Concepts of the Nature of Mind*, London, 1965. The wealth of cross-disciplinary confusions over mental images is displayed in both volumes, which both include papers by philosophers, psychologists and neurophysiologists. Neither editor seems to think that much of what he presents is a dead horse, which strengthens my occasionally flagging conviction that I am not beating one. On the other hand there are scientists who have expressed clear and explicit rejections of imagistic confusions. See, e.g., G. W. Zopf,

‘Sensory Homeostasis’ in Wiener and Schadé, *op. cit.* , esp. p. 118, and D. M. MacKay,

‘Internal Representation of the External World’, unpublished, read at the Avionics Panel Symposium on Natural and Artificial Logic Processors, Athens, July 15–19, 1963.

2 H. B. Barlow, ‘Possible Principles Underlying the Transformations of Sensory Messages’ in *Sensory Communication* (*op. cit.*) offers a particularly insightful account of the ‘editorial’

function of afferent neural activity and the depletion of information that is the necessary concomitant of such analysis.

3 J. M. Shorter, ‘Imagination’, *Mind*, LXI, 1952, pp. 528–42.

4 Counter-examples spring to mind, but are they really counterexamples? All the ones that have so far occurred to me turn out on reflection to be cases of imagining myself seeing –

with the aid of large mirrors – the inside and outside of the barn, imagining a (partially) transparent barn, imagining looking in the windows and so forth. These are all from a point of view in the sense I mean. A written description, however, is not bound by these limitations; from what point of view is the description: ‘the barn is dark red with black rafters and a pine floor’?

5 In the unusual phenomenon of ‘eidetic imagery’, the subject *can* read off or count off the details of his ‘memory image’, and this may seem to provide the fatal counter-example to this view. (See G. Allport, ‘Eidetic Imagery’, *British Journal of Psychology*, XV, 1924, pp.

99–120.) Yet the fact that such ‘eidetic memory images’ actually appear to be projected or superimposed on the subject’s normal visual field (so that if the subject shifts his gaze the position of the memory image in his visual field remains fixed, and ‘moves with the eye’) strongly suggests that in these cases the actual image of retinal stimulation is somehow retained at or very near the retina and superimposed on incoming stimulation. In these

rare cases, then, the memory mechanism must operate *prior to* afferent analysis, at a time when there still is a physical image.

6 Penfield, *op. cit.* Some of Penfield's interpretations of his results have been widely criticized, but the results themselves are remarkable. It would be expected that hallucinations would have to be the exception rather than the rule in the brain for event-types to *acquire* content in the first place, and this is in fact supported by evidence.

Amputees usually experience 'phantom limb' sensations that seem to come from the missing limb; an amputee may feel that he not only still has the leg, but that it is itching or hot or bent at the knee. These phenomena, which occur off and on for years following amputation, are nearly universal in amputees, with one interesting exception. In cases where the amputation occurred in infancy, before the child developed the use and coordination of the limb, phantom limb is rarely experienced, and in cases where

amputation occurred just after birth, no phantom limb is ever experienced (see M. Simmel,

'Phantom Experiences following Amputation in Childhood', *Journal of Neurology, Neurosurgery and Psychiatry,* XXV, 1962, pp. 69–78).

7 Other phenomena less well known to philosophers also favour a descriptional explanation.

See, e.g., W. R. Brain's account of the reports of patients who have their sight surgically restored, in 'Some Reflections on Mind and Brain', *Brain*, LXXXVI, 1963, p. 381; the controversial accounts of newly sighted adults' efforts to learn to see, in M. von Senden, *Raum- und Gestaltauffassung bei operierten Blindgeborenen vor und nach der Operation*, Leipzig, 1932, translated with appendices by P. Heath as *Space and Sight, the Perception of Space and Shape in the congenitally blind before and after operation*, London, 1960; I.

Kohler's experiments with inverting spectacles (a good account of these and similar experiments is found in J. G. Taylor, *The Behavioral Basis of Perception*, New Haven, 1962); and the disorder called simultanagnosia, M.

Kinsbourne and E. K. Warrington, ‘A Disorder of Simultaneous Form Perception’, *Brain*, LXXXV, 1962, pp. 461–86 and A. R.

Luria, *et al.* , ‘Disorders of Ocular Movement in a Case of Simultanagnosia’, *Brain*, LXXXVI, 1963, pp. 219–28.

8 Cf. Wittgenstein, ‘But the existence of this feeling of strangeness does not give us a reason for saying that every object we know well and which does not seem strange to us gives us a feeling of familiarity’, *op. cit.* , i. 596. See also i. 597, i. 605.

9 Muntz, *op. cit.* and Wooldridge, *op. cit.* , pp. 46–50.

10 Having found no room for images in the sub-personal account of perception, we can say that ‘mental image’ and its kin are poor candidates for referring expressions in science; having found further that nothing with the traits of genuine images is to be found at the personal level either allows us to conclude that ‘mental image’ is valueless as a referring expression under *any* circumstances.

11 J. J. C. Smart discusses this and other colour phenomena in *Philosophy and Scientific Realism*, London, 1963, Ch. IV and cites (with some philosophically negligible distortion) the particularly remarkable findings of E. H. Land in ‘Experiments in Color Vision’, *Scientific American*, No. 200 (May) 1959, pp. 84–99.

12 But not infinite. It is easy to make a device (a colour discriminator) that gives the same output (produces the same colour experience) when given a number of different inputs with no distinguishing feature in common. Suppose we choose three such inputs. Then we simply make three receptors each sensitive to just one of these and wire them disjunctively to the output, so that any *one* of them firing causes the output to fire. A single receptor can be sensitive only to conditions having some physical similarity (unless it in turn branches disjunctively into a number of receptors), so an infinite number of different conditions which could not be sorted into a finite number of families of conditions would require an infinite number of different receptors in the discrimination system.

13 D. M. Armstrong has advanced this position in discussion with me.

8

THINKING AND REASONING

1 Cf. S. Munsat, ‘What is a Process?’, *American Philosophical Quarterly*, 1969, pp. 79–83.

2 Ryle, *op. cit.* , p. 286ff.

3 *Ibid.* , p. 285.

4 Striking cases of confusion of level often occur in the casual talk of neurologists and psychophysicists. Once a psychophysicist, in explaining to me his interpretation of certain data concerning afferent analysis in visual perception said of his human subject, ‘he performs a statistical analysis on the incoming information, making decisions with regard to what is mere noise and what is significant, by trying to maintain a low false-positive rate.’

The *person*, of course, does nothing of the kind; the person in this instance was just looking for spots on pieces of paper.

5 J. Bennett, *Rationality*, London, 1964, p. 117.

6 See Feigenbaum and Feldman’s anthology, *op. cit.* , esp. Newell and Simon, *op. cit.* See also for a critical view, H. Dreyfus, ‘Alchemy and Artificial Intelligence’, *RAND Memo*, P 3244, Dec. 1965, and my rebuttal, ‘Machine Traces and Protocol Statements’.

7 Cf. A. M. Turing’s famous paper, ‘Computing Machines and Intelligence’, sec. 7, in *Mind*, LIX, 1950, pp. 433–60, reprinted in Anderson, *Minds and Machines*.

8 Anscombe, *Intention*, p. 80.

9 D. C. Davidson, ‘Actions, Reasons and Causes’, *Journal of Philosophy*, Vol. LX, No. 23

(Nov. 1963), pp. 685–700.

10 Anscombe, *op. cit.* , p. 15.

11 *Ibid.* , pp. 23–4.

9

ACTIONS AND INTENTIONS

1 Anscombe, *op. cit.* , pp. 24–5.

2 *Ibid.* , p. 82.

3 *Ibid.* , p. 82.

4 *Ibid.* , p. 57.

5 *Ibid.* , p. 25.

6 This account of intentional actions carries the corollary that that in virtue of which a particular motion is a particular intentional action is in principle – if not in practice –

physically determinable in some centralist theory. Although the account here is in substantial agreement with Anscombe's analysis, she has an argument (pp. 28–9) purporting to disprove this corollary, to establish the unbridgeable Intentionalist gap for intentional actions. For a rebuttal, see my 'Features of Intentional Actions', *Philosophy and Phenomenological Research*, XXIX, 1968, pp. 232–44.

7 Anscombe, *op. cit.* , pp. 48–9.

8 *Ibid.* , p. 52.

9 D. M. MacKay points out that the relatively macroscopic size of neuronal events and the redundancy requirements noted in Chapter 3 rule out any cumulative effect on behaviour of quantum level randomness ('Brain and Will', *Listener, 57* (1957), pp. 788–9).

10 Anscombe, *op. cit.* , p. 24.

11 *Ibid.* , p. 34.

10

LANGUAGE AND UNDERSTANDING

1 Cf. M. Scriven in D. M. Mackay *et al.* , ‘Computers and Comprehension’ (*op. cit.*): ‘One either knows three threes are nine or one doesn’t. It isn’t the kind of thing one is said to understand’ (p. 33).

2 E. Edwards, *Information Transmission*, London, 1964, p. 39.

3 D. M. MacKay, ‘Linguistic and Non-Linguistic “Understanding” of Linguistic Tokens’, p.

42. The Intentionality of intentions forces a revision of this if we are to be rigorous. Few people, if any, would ever intend an utterance to have a selective function on the range of possible states of the receiver (‘I intended no such thing. I merely was trying to tell him something!’). We can say that the effect which is the necessary condition of his intention being fulfilled is this selective function. I do not want to suggest in any case that this is *the* definition of the meaning of an utterance, but only that it is an illuminating one.

4 *Ibid.* , p. 9.

INDEX

access *see* introspection

act/action 22, 37–8;

intentional 39, 43, 147, 184–200;

mental 147, 214

activity 195–6, 212;

personal 103–4, 166;

reasons 187, 191, 193

adaptiveness *see* appropriateness

afferent 51, 58;

-efferent connection 52, 54–7, 62–5, 68–70, 76–9, 81, 83–7, 91, 93, 102, 106, 186–7; ascription 82, 92, 94;

awareness 137;

introspection 124;

mental imagery 154;

thinking 173

Allport, G. 228

ambiguity 9, 60, 62, 79, 83, 90, 143, 177

analysis/analyticity 8, 14–16;

actions 184;

ascription 81, 83, 86, 92–3, 102–3, 105–6; awareness 134, 137, 144–5, 147;

evolution 51, 56, 59; imagery 151–2, 154, 159–60, 168–9; intentionality 27–8;

introspection 111, 124–5;

language 214–15;

thinking 168–9, 177

Anscombe, G.E.M. 112–13, 174, 180–1, 184–5, 187–90, 192, 195, 199–200, 220, 225, 227,

230–1

Anselm 23

aphasia 74, 120

appropriateness 40–1, 44;

ascription 80, 90, 96;

evolution 50–1, 54, 62–4, 75; thinking 171

Arbib, M. 61, 223

Aristotle 78, 174–5

Armstrong, D.M. 229

attention 45, 65, 111, 132–3, 135, 137, 139–40, 157, 185

avowal 112, 225

awareness 128–49, 154–5; intentions 184–9, 191–2, 194–8, 200;

reasoning 164, 168, 171, 173, 176, 180

Barlow, H.B. 227

baroque 64, 69

behaviourism 25, 35–7, 40–2, 45, 69, 82–3, 101

belief 23, 27;

ascription 81, 86–8, 91, 100, 107;

goals 78;

intentions 34, 37, 40, 44, 184, 187, 191, 201–3; mental imagery 159;

reasoning 164;

understanding 201–3

Bennett, J. 167, 230

Bergmann, G. 226

Berkeley, G. 3

bracketing *see* epoché

brain 46–7, 51–2;

ascription 80–4, 86, 88–9, 92–5, 97–8, 101–6, 108;

consciousness 136–7, 139–41, 144, 146;

goals 70–1, 77–9;

intentions 193–6, 200;

introspection 111–12, 118–20, 125–7; mental imagery 150–2, 155;

reasoning 164, 167, 170, 173–5; structures 55–7, 59–68; understanding 213–15

Brain, W.R. 228

Brentano, F. 21–5, 29, 31, 35, 220

brute force computing 169–70

Carnap, R. 221

category 7–8, 12, 14, 45, 107, 218

cause 4, 8;

ascription 85, 88, 91–2, 94;

goals 70;

intentions 37–9, 193, 195, 197–8; introspection 116, 119–20;

mental imagery 155;

reasoning 175, 179–83;

structures 51, 53–4, 61; understanding 207, 209

centralism 45, 88, 92, 95

cerebellum 140

certainty *see* introspection

Chisholm, R. 23–4, 26, 29, 34, 220–1

Chomsky, N. 117, 225

colour 61, 104, 147, 149–50, 152–3, 157–63

commands 92, 117–18, 120, 186, 189–92, 198, 201

compulsion of pain behaviour 102, 106

computers 43–4, 48–9;

ascription 88, 100;

goals 72–3;

introspection 114, 116, 121–3; reasoning 167, 169–70, 172–3;
understanding 204–5

conceive versus imagine 155

consciousness 4, 45;

awareness 128–30, 133, 142–3, 146–8; intentions 194;

introspection 111–12;

mental imagery 149, 156, 158–9; reasoning 164–6;

understanding 215

content 21, 23;

ascription 80–1, 85–92, 94–9; awareness 131, 133, 136, 138, 146, 148;

goals 79;

information 47;

intentions 29, 31, 43–4, 186–9, 192, 199;

introspection 112, 117–21, 126–7; mental imagery 154–5, 157;

reasoning 164, 183;

structures 60, 62;

understanding 201, 203, 206, 210–14

context 7, 9–11;

ascription 90, 92, 103;

awareness 133;

intentions 26, 185;

ontology 14–15, 17;

understanding 208, 210

control of behaviour 50–7, 64–79;

ascription 80–9, 91, 93–6, 98–102, 106;

awareness 131–4, 136, 139–41, 143, 147;

intentions 191, 195–6, 198–200;

introspection 112, 117–19, 121;

mental imagery 158;

reasoning 167, 171–3, 178–80, 182;

understanding 203, 205

cortex 140, 155

criterion of awareness 128–49, 154–5; intentions 184–9, 191–2, 194–8, 200;

reasoning 164, 168, 171, 173, 176, 180

cybernetics 13, 43, 135

data language 25

Davidson, D.C. 230

deprivation experiments 67, 69

Descartes, R. 3–4, 112, 155, 218

description 23–4, 30–1;

ascription 80, 85, 87, 89, 95–6, 99, 103;

goals 73;

information 48;

intentions 43–4, 197, 199, 205;

introspection 115, 120, 122;

mental images as 126, 132, 153–5; motions 184–6, 188, 191–2

desire 6, 8, 14, 23, 37, 81, 87–8, 95, 100, 129, 184

direct quotation (*oratio recta*) 28–9, 33, 92, 118

discrimination 61, 65–8, 70, 80, 82–4, 103–4, 106, 156, 161–2

disposition 9–10, 188, 214

dreams 90, 144, 154

Dreyfus, H. 230

dualism 3–4, 6, 14–15, 99

Edwards, E. 231

efferent 117, 120, 173

eidetic imagery 228

emergent phenomena or qualities 101, 107, 159, 163

enthymeme 174

environment 10, 26;

ascription 80–2, 84, 86–7, 89, 93;

awareness 130, 139–40;

goals 71, 73, 79;

information 50;

intentions 40, 43–4, 50, 184;

introspection 124;

mental imagery 152–3, 157;

structures 51–2, 54–5, 63–5, 67–8

epiphenomena 4, 101

epistemology 26–7, 83, 93, 189

epoché 26

evolution 44, 90;

brain 46;

content 80;

goals 70–1, 76, 78–9; pain 102;

structures 51–69

existence 96, 98, 102, 159, 168;

content 85;

evolution 67, 71;

identity 6–19;

inexistence 21–3;

intention 25–7, 35, 44

explanation 10, 17–19;

ascription 80, 82, 87, 89, 95–7, 99–100;

awareness 128, 134;

intentions 21, 37–41, 184, 193, 197, 200;

introspection 113–14, 126;

mental imagery 152, 159;

pain 101–8;

reasoning 165, 167–8, 178, 180;

structures 52

expression 17, 23–6;

ascription 85, 92, 99, 101;

awareness 133, 141;

goals 72;

intentions 28, 30, 32, 42, 45, 188;

introspection 112, 123–7;

reasoning 175;

understanding 207

extension 16, 31–5;

ascription 80, 85, 87–90, 95–7, 99–100;

intentions 37–8, 40, 42–3, 45–8, 197;

understanding 214

facts 69, 152, 162, 174, 193, 201–2, 206–8, 215

feedback 56, 69, 123, 137, 139, 180, 188

Feigenbaum, E.A. 223, 230

Feigl, H. 217

Feldman, J. 223, 230

Fraassen, B. van 218–19

free will 198

Freud, S. 119, 144–5, 222

function 7, 10;

ascription 80, 83, 88–9, 91–2, 94, 96–7, 101;

awareness 129, 136–7, 142, 146;

goals 70;

intentions 32, 189, 195, 198;

introspection 114–17, 124, 126;

mental imagery 150, 163;

structures 51–2, 56–7, 60–5, 69; understanding 203, 209–10, 212, 214–15

fusion 15–17, 107

Geach, P.T. 27, 220–1

generalization of learning 36, 77

glial cells 57

goals/goal-directed behaviour 36, 43–4, 48, 70–9, 81–2, 89, 100, 118

GPS (General Problem Solver) 72–3, 76

grammar 9, 11, 119

hallucinations 123–4, 149, 155

Hamlyn, D. 227

Harlow, H.F. 64

hierarchy 72, 78–9, 117, 119, 122, 137

Hirst, R.J. 227

Hobbes, T. 3

Hubel, D.A. 223–4

Hume, D. 37–9

Husserl, E. 26

hypnotism 123, 144, 200

idea 14, 20, 89–90, 132, 159, 177, 183, 193, 210

idealism 3–4

identity 6–20, 28;

ascription 101, 107;

intentions 30–1, 40;

introspection 118, 120–1, 124, 126;

understanding 211

illusions 25, 124, 161–2, 176, 201

images 5, 20, 65, 121–2, 147–63, 168, 215

imagination 28, 91, 152, 156, 167–8

indirect quotation (*oratio obliqua*) 33, 92, 118

infallibility *see* introspection

inferential and non-inferential knowledge *see* know

information 32–3, 43–5;

ascription 81, 86, 91–4, 97–8, 100;

awareness 132, 140–1, 143–4; goals 78;

intelligent uses of 46–50;

intentions 189, 191, 193–5, 198–9; language 209–13;

mental imagery 151, 156, 158, 163;

reasoning 168, 170–6, 179, 182–3; structures 51, 55, 57, 60–2, 65; understanding 201, 203–8

inhibition 59, 64, 93, 120

insightful behaviour 73

instinctual behaviour 55, 66–7, 172

intelligence 44, 49–51;

ascription 81, 86, 91–2, 94;

goals 74–5;

structures 56, 62, 66; understanding 204, 214;

use of information 46–50

intension 31–2, 85

intentionality 20–48, 80–1; actions 184–200;

ascription 87–90, 92, 95–101;

awareness 129–30, 142, 147;

intentions 185;

introspection 121, 125–6;

mental imagery 150;

reasoning 174, 182–3;

understanding 201, 209, 213–14

interaction 4, 14, 87, 89

introspection 82, 101;

awareness 128, 131–2, 134, 144;

certainty 111–27;

reasoning 164–5, 168–70, 173, 175

intuition 7, 11;

ascription 81;

awareness 135, 141, 146;

mental imagery 149–57;

reasoning 170, 173;

understanding 211

Kinsbourne, M. 229

know 12, 24–7;

ascription 82–3, 88, 91, 94, 102–3; awareness 131, 133–5, 143;

goals 74–5;

information 49;

intentions 36, 44, 186–91, 196–7; mental imagery 156–7, 159–61, 163;

reasoning 164, 167–8, 170–1, 173–4, 176–83; structures 55, 66;

understanding 201–15

know. introspection 111–14, 116

Kohler, I. 228

Körner, S. 220–1

Lamarck, J. 66

Lambert, K. 218–19

Land, E.H. 229

language 8, 17–23;

ascription 101;

awareness 129, 140–1, 144–8; content 92–100;

information 209–13;

introspection 117–19, 123, 126–7;

mental 25, 30, 35–6, 42, 46, 75, 80, 108, 113;

reasoning 169–70, 172–3, 177;

understanding 201–15

Lashley, K.S. 36, 144, 164, 221, 225–6

learning 36, 50, 56, 65–71, 76, 78, 122, 139–40, 181, 208

Lenneberg, E. 225

Leonard, H. 218

Lettvin, J.Y. 223–4

location of pain *see* pain

Locke, J. 159

locomotion *see* motion

logical 6–8, 13;

ascription 81, 83, 85, 96;

awareness 139;

information 46;

intentions 22;

introspection 114–17, 122, 124, 126;

ontology 15–17;

reasoning 165–7, 172–7;

understanding 214

Luria, A.R. 229

McGuinness, F. 224

machines 6, 98;

awareness 135–8, 142, 147–8; introspection 112, 114–16;

perceiving 117–27;

reasoning 169–70, 173

Mackay, D.M. 50, 212–13, 223–4, 227, 231

man 8, 33;

ascription 98, 102, 106;

awareness 130–4, 139–40, 144;

goals 74, 78;

intentions 187–9, 195–6, 200;

introspection 111, 121–5;

mental imagery 151–3, 161;

reasoning 175–6;

structures 56, 69;

understanding 208, 210, 214

materialism 3, 10

meaning 9, 13;

ascription 82, 84, 87, 92, 94, 99;

awareness 129, 132, 136;

goals 79;

information 47;

intentions 21, 31, 37, 195–6; mental imagery 159, 162;

reasoning 176;

structures 57, 60;

understanding 205, 212

mechanical 102–3, 106–7, 145–6, 198

Melden, A.I. 222

memory 49, 87, 97, 157, 178

mental act *see* act

mental language *see* language

mental process *see* process

message 43, 47;

ascription 87–9, 92, 98–9; intentions 194–6;

reasoning 180, 194–6;

structures 51, 61;

understanding 209–10, 214

Meyer, R.K. 219

mind 22–3, 26–7;

ascription 80, 98, 100–1, 106–8; awareness 128, 134, 138, 141, 144;

imagery 149–63;

intentions 29, 42, 45–6, 195–6; introspection 111–12;

language 3–108;

ontological problem 3–19;

reasoning 165, 169, 173, 177;

science 3–6;

understanding 212–13

modal logic 28

mongrel intentional expressions 26–7, 30, 34

monism 6

motion 4, 35–7;

ascription 81–2, 85, 88, 105–6; awareness 131–2, 138–9;

goals 70, 75;

intentions 43, 184–9, 191, 197–8, 200;

mental imagery 158;

reasoning 175, 180;

structures 51, 55, 63, 65, 68

motive 5, 88, 129

Munsat, S. 229

Muntz, W.R.A. 224, 229

Nagel, T. 217, 219

natural selection *see* evolution

Neisser, U. 226

neural net 77, 92, 101, 121

neural structures 63, 71, 80, 92

neurons 52–3, 57–62, 77, 85, 91, 155

Newell, A. 72, 223, 230

object, intentional *see* intentionality

observation 25, 54;

awareness 134;

goals 71;

intentions 186–7, 189;

introspection 112, 119;

mental imagery 156;

reasoning 165;

understanding 212

occasion 4, 34;

ascription 97, 100, 103, 107;

awareness 135–6, 139;

information 49;

intentions 195–6;

reasoning 165, 174, 179, 181;

structures 60, 64

ontology 3–19, 145

operant behaviour 56–7

operation 14, 43;

ascription 88, 90, 99, 102;

awareness 141;

introspection 123;

mental imagery 157;

reasoning 166–70, 172–5;

structures 61, 63

oratio obliqua see indirect quotation

oratio recta see direct quotation

ordinary language *see* language

ostension of qualities 104

pain 6, 8;

ascription 82, 86, 92;

awareness 141, 145;

introspection 112–14, 116;

levels 101–8;

ontology 14, 18;

reasoning 176, 181;

structures 55, 67, 69–70

para-mechanical hypotheses *see* mechanical

parallelism 4, 13–14, 17, 26, 108, 113, 173–4

pattern recognition 55, 77, 83

Pavlov, I.P. 182

Penfield, W. 120, 225, 227–8

perception 40, 47, 122–4, 150–6, 158, 162–3

peripheralism 45–7, 60, 71, 76–7, 81, 101, 157

person 18, 71;

ascription 96, 101–8;

awareness 130, 134, 137–9, 141, 143–4, 147;

intentions 184, 191, 194, 196, 198–200;

introspection 111, 113–14, 125–7; mental imagery 149, 151–5, 159–60; reasoning 165–7, 171–2, 175, 177–82; understanding 201–5, 208, 211–15

phantom limb 228

phenomenology 26, 35, 37, 101

phoneme 117, 207

physicalism 4, 10, 13–15, 120

Pittendrigh, C.S. 41, 222

Place U.T. 217

pondering 167–8, 179

practical knowledge *see* know

pre-wiring 55, 57, 62, 65, 67–9

precision of content ascription 80–108, 117, 126, 186, 201

prediction 87, 96

primary qualities *see* quality

print-out 117, 123, 169–70

private qualities *see* quality

process 6, 10;

ascription 86, 97–8, 100–2, 104–6; awareness 144, 146, 148;

goals 77;

intentions 20, 187, 191, 194;

introspection 116, 118, 120, 126;

mental imagery 151–2, 160;

people 164–75;

reasoning 176, 179–80, 183;

structures 51, 54, 56, 59–60, 65–6; understanding 213

proposition 21, 23–4;

ascription 92;

awareness 133;

intentions 28–31, 33;

reasoning 165;

understanding 202–3, 206–7, 213

proprioception 180, 186, 189–90

protocol 119, 169–71, 175, 178

psychoanalysis 177, 179

punishment 69–71, 184

purpose 34, 43–4, 59, 62, 78, 122, 144, 165, 181

Putnam, H. 18, 114–17

quality 65, 103–4, 106, 150, 158–60, 163

quantifier 218–19

Quine, W.V.O. 7–8, 24, 27–8, 32–5

quotation *see* direct quotation;

indirect quotation

rationality/rationalizing 96, 179, 191

reason/reasoning 45, 54;

ascription 90–1;

awareness 132–3;

goals 70–2, 74–5;

intentions 187, 190–2, 194, 197, 199–200;

introspection 111, 122;

mental imagery 161;

structures 56;

thinking 164–83;

understanding 203, 214–15

recognition 6, 54;

ascription 83, 107;

awareness 142, 145;

goals 71, 77;

reasoning 178;

structures 68–9

reduction 33, 42, 45, 48

redundancy 60, 62, 123

reference 16–18, 21, 29, 31, 85, 93, 100, 160, 211

reflex 55–6, 67, 79, 101, 138, 143–4, 180

reliability 61, 209

report 47, 85;

ascription 88, 90–2;

awareness 128, 131–4, 142, 146;

intentions 187–9, 191, 199;

introspection 112–14, 116–17, 121–7; mental imagery 155–7, 162;

reasoning 165, 168–9, 176–9, 181;

understanding 201, 206–7

representation 150, 152

response 36–7, 45;

ascription 82–3, 85, 87, 93, 101, 106;

awareness 133, 135, 138, 143;

goals 70–1, 79;

information 49–50;

introspection 122;

mental imagery 158;

reasoning 177, 182;

structures 51, 53–6, 63; understanding 208

responsibility 20, 119, 125, 175, 184, 198, 200

reward 69–71

Ryle, G. 6, 8, 14, 98, 106, 111–13, 120, 145–6, 166, 207

Sayre, K.M. 142, 226

Scriven, M. 142, 226, 231

secondary qualities *see* quality

seeing 93, 125, 143, 153–6, 161, 195

seem 4, 6–10;

ascription 82, 86, 93–5; awareness 132, 134, 136, 140, 143–4; goals 73–5;

intentions 26, 33, 43, 190, 193;

introspection 111–12, 124–5;

mental imagery 155, 160, 162;

ontology 14, 16;

reasoning 164, 176, 178, 180;

structures 59, 63, 67–8; understanding 201, 206–8, 212

semantics 8, 16–17, 93, 145

Senden, M. von 228

sensation 5, 76, 102–5, 139, 186

sense-data 147

sentence 9, 11;

ascription 97–9;

awareness 136, 138, 145–6; information 46;

intentions 20–5, 28–34, 38, 42, 45;

introspection 117–18, 120, 122–4, 126;

ontology 13, 15–19;

understanding 204, 206, 208–12

Sheffler, I. 222

Shorter, J.M. 152, 227

Shwayder, D. 222

significance 15, 43–4;

ascription 80, 83–4;

goals 79;

intentions 190;

introspection 122;

reasoning 180, 182;

structures 51–2, 54–6, 62, 68

Simmel, M. 228

Simon, H.A. 72, 223, 230

Skinner, B.F. 56–7, 122, 223, 225

Smart, J.J.C. 217, 219, 229

Smythies, J.R. 227

Sommers, F. 217

speech 33, 35;

ascription 98;

awareness 128–9, 133, 136–8, 140, 144;

introspection 117–18, 121–5;

reasoning 168

Stampe, D.W. 225

state 10, 18–19;

ascription 80–2, 85–7, 89, 91–2, 94–6, 98–9, 107;

awareness 133, 143;

goals 72–4, 76–7, 79;

information 47, 50;

intentions 20, 26–7, 34, 41, 43–5, 186–8, 195–6, 198;

introspection 114–17, 121–2, 124, 126–7; mental imagery 163;

reasoning 164, 169;

structures 67;

understanding 201–5, 209, 211–14

stimulus 36, 45;

ascription 82–6, 90–4, 101;

awareness 138;

goals 70–1, 76–7;

information 50;

introspection 112, 119, 121;

structures 51–3, 55–7, 60, 63, 65, 67, 69; understanding 209

structures 71–2, 78;

ascription 80–1, 87, 89, 91–2, 94, 97;

awareness 137;

evolution 51–70;

mental imagery 160–1, 163

sub-personal level of explanation *see* person

subjectivity 163

syllogism 174

synapses 58, 60, 87

syncategorematic 218

syntactic 12, 97–8, 119, 136, 209

Taylor, C. 43, 220–3, 227

Taylor, J.G. 228

teleology 42, 222–3

Theophrastus 187–8

thinking 18–19, 30;

awareness 139, 148;

intentions 191, 193, 195–6; introspection 112, 120, 127;

mental imagery 149, 155;

reasoning 164–83

thought 5–6, 8;

ascription 107;

awareness 140, 144–6, 148;

information 47;

intentions 20, 23, 191, 193, 195–6; introspection 112, 114, 118–21, 126–7; ontology 18–19;

reasoning 164–5, 172, 174, 176–8, 182–3; understanding 213

threshold 58–9, 62, 156

token 97, 210

Tolstoy, L. 153

translation 30

trial and error 69–70, 167, 169

tropism 55–6, 79, 172

Turing, A. 114–16, 126, 230

type 12, 40;

ascription 88, 90–1, 93, 97, 99;

information 48–9;

introspection 120;

mental imagery 161;

structures 57

unconscious 129, 136, 142–6, 170, 210

understanding 5, 13, 88–9, 97, 121, 181, 201–15

unlearning 65

Van Fraassen *see* Fraassen

variables 59, 61–2, 78, 89

verbal behaviour 7, 10;

ascription 92, 98–9;

awareness 140;

information 49;

intentions 35, 188–9, 194, 198–200;

introspection 112, 116, 118–19, 122–6; reasoning 172, 176–7, 182;

understanding 201, 204–5, 209, 214

vitalism 60, 78, 87, 89

volition 192, 195–6, 215

voluntary 102, 190, 197, 200

Von Senden *see* Senden

wanting 22, 30, 95–6

Warrington, E.K. 229

Wiesel, T.N. 223

will power 102

willing 16, 179, 192–7

Wittgenstein, L. 106–7, 112–13, 154, 160, 224–5, 229

Wooldridge, D. 223, 229

Zeman, J. 97, 224

Zopf, G.W. 227

Document Outline

- CONTENTS
- PREFACE TO THE ROUTLEDGE CLASSICS EDITION
- PREFACE TO SECOND EDITION
- PREFACE TO FIRST EDITION
- Part I The Language of Mind
  - 1 THE ONTOLOGICAL PROBLEM OF MIND
    - I THE MIND AND SCIENCE
    - II EXISTENCE AND IDENTITY
  - 2 INTENTIONALITY
    - III THE PROBLEM OF INTENTIONALITY
    - IV TWO BLIND ALLEYS
    - V THE WAY OUT
  - 3 EVOLUTION IN THE BRAIN
    - VI THE INTELLIGENT USE OF INFORMATION
    - VII THE EVOLUTION OF APPROPRIATE STRUCTURES
    - VIII GOAL-DIRECTED BEHAVIOUR
  - 4 THE ASCRIPTION OF CONTENT
    - IX FUNCTION AND CONTENT
    - X LANGUAGE AND CONTENT
    - XI PERSONAL AND SUB-PERSONAL LEVELS OF EXPLANATION: PAIN
- Part II Consciousness
  - 5 INTROSPECTIVE CERTAINTY
    - XII THE CERTAINTY OF CERTAIN UTTERANCES
    - XIII A PERCEIVING MACHINE
  - 6 AWARENESS AND CONSCIOUSNESS
    - XIV THE ORDINARY WORDS
    - XV AWARENESS AND CONTROL
    - XVI CONSCIOUSNESS
  - 7 MENTAL IMAGERY
    - XVII THE NATURE OF IMAGES AND THE INTROSPECTIVE TRAP

        - XVIII COLOURS
    - 8 THINKING AND REASONING
        - XIX PEOPLE AND PROCESSES
        - XX REASONS AND CAUSES
    - 9 ACTIONS AND INTENTIONS
        - XXI INTENTIONAL ACTIONS
        - XXII WILLING
        - XXIII THE IMPORTANCE OF INTENTIONAL ACTIONS
    - 10 LANGUAGE AND UNDERSTANDING
        - XXIV KNOWING AND UNDERSTANDING
        - XXV LANGUAGE AND INFORMATION
        - XXVI CONCLUSIONS
- NOTES
- INDEX